Kapazität zur Bildung von Sekundärstrukturen bei Ribonucleinsäuren

Kurt Stüber

Kurt Stüber

KAPAZITÄT ZUR BILDUNG VON SEKUNDÄRSTRUKTUREN BEI RIBONUCLEINSÄUREN

Computeruntersuchung
speziell bei den Bakteriophagen
MS2 und φX174

1978

Aus dem botanischen Institut der Universität Bonn

KAPAZITÄT ZUR BILDUNG VON SEKUNDÄRSTRUKTUREN BEI RIBONUCLEINSÄUREN

- - - -

Computeruntersuchung
speziell bei den Bakteriophagen
MS2 und φX174

Wissenschaftliche Arbeit
im Rahmen der Diplom-Hauptprüfung
vorgelegt von

Kurt Stüber*

1978

*Adresse: Starenweg 7, 5300 Bonn 1

Angefertigt im
Institut für Botanik
der Universität Bonn

Referent:
Prof. Dr. Dieter Klämbt

Koreferent:
Prof. Dr. Jobst-Heinrich
Klemme

Datum der Abgabe:
7.4.1978

Gewidmet in Dankbarkeit
meinem zukünftigen
Schwiegervater

INHALTSVERZEICHNIS

0 ABKÜRZUNGSVERZEICHNIS

A	Adenin
A_1	Nummer des dem 5′-Ende zu gelegenen Nucleotids des 1. Basenpaares einer Helix
A_2	Nummer des dem 3′-Ende zu gelegenen Nucleotids des 1. Basenpaares einer Helix
BL	Bulge loop = Ausbauchungs- oder Bauchungsschleife
C	Cytosin
C $C((N \cdot N)_i)$ C-Menge	Menge der zu einem Basenpaar $(N \cdot N)_i$ kompatiblen Basenpaare oder auch der zu einer Faltung kompatiblen Basenpaare
e	Euler′sche Zahl = 2.7182818284...
E_1	Nummer des dem 5′-Ende zu gelegenen Nucleotids des letzten Basenpaares einer Helix
E_2	Nummer des dem 3′-Ende zu gelegenen Nucleotids des letzten Basenpaares einer Helix
E(A)	Erwartungswert der Anzahl der Adenin in einer Nucleinsäure
E(C)	Erwartungswert der Anzahl der Cytosin in einer Nucleinsäure
E(G)	Erwartungswert der Anzahl der Guanin in einer Nucleinsäure
$E_{abs}(H^n)$	Erwartungswert der absoluten Anzahl möglicher Helices der Länge n in einer Ribonucleinsäure
$E_{int}(H^n)$	Erwartungswert der integralen Anzahl möglicher Helices der Länge n in einer Ribonucleinsäure
E(N)	Erwartungswert der Anzahl der Base N in einer Nucleinsäure, statt N kann auch U,C,A oder G stehen
E(U)	Erwartungswert der Anzahl der Uracil in einer Nucleinsäure
G	Guanin
H^n	Helix der Länge n

H_n	n-te Helix
H_m^n	m-te Helix der Länge n
HP	Hairpin loop = Haarnadelschleife
i	Variabler Index
I	Informationskapazität
I(N→)	Informationskapazität der Transitionen von der Base N auf die folgende, statt N kann auch U,C,A oder G stehen
I(→)	Informationskapazität errechnet aus den Transitionswahrscheinlichkeiten
IL	Internal loop = innere Schleife
j	Variabler Index
k	Variabler Index
kcal	Kilokalorien
kJ	Kilojoule
$K^n(M)$	Anzahl der sterisch möglichen Kombinationen (von Teilsequenzen) der Länge n in einer Ribonucleinsäure der Länge M
l	Variabler Index
L	Anzahl verschiedener in eine Faltung aufnehmbarer Basenpaare
ld	Logarithmus zur Basis 2 = Logarithmus dualis
m	Variabler Index
M	Anzahl der Basen bzw. Nucleotide einer RNS = Länge einer RNS
mRNS	Messenger-Ribonucleinsäure
MS2	Coliphage MS2
n	Variabler Index
N	Nucleotid, steht für die Basen Adenin, Cytosin,Guanin oder Uracil
N-N	2er Sequenz, statt N kann A,C,G oder U stehen
N-N-N	3er Sequenz, statt N kann A,C,G oder U stehen, entsprechend werden auch längere Sequenzen geschrieben: N-N-N-N, N-N-N-N-N-N, etc
(N·N)	Basenpaar, statt N kann A,C,G oder U stehen.

$(N_i \cdot N_j)$	Basenpaar zwischen den Nucleotiden mit den Nummern i und j ; statt N kann A,C,G oder U stehen
$(N \cdot N)_i$	Basenpaar mit der Nummer i; statt N kann A,C,G oder U stehen
n_{max}	grösste mögliche Länge einer Helix in einer RNS
NS	Nucleinsäure
pA	Wahrscheinlichkeit für das Auftreten von Adenin an einer beliebigen Stelle einer Nucleinsäure
pC	Wahrscheinlichkeit für das Auftreten von Cytosin an einer beliebigen Stelle einer Nucleinsäure
pF	Faltungskapazität = Wahrscheinlichkeit der Bildung eines Basenpaares bei Zusammentreffen zweier beliebiger Nucleotide einer Nucleinsäure
pG	Wahrscheinlichkeit für das Auftreten von Guanin an einer beliebigen Stelle einer Nucleinsäure
$p_{abs}(H^n)$	Absolute Wahrscheinlichkeit für das Auftreten einer Helix der Länge n bei beliebigem Kontakt zweier Teilsequenzen der Länge n einer Ribonucleinsäure
$p_{int}(H^n)$	Integrale Wahrscheinlichkeit für das Auftreten einer Helix der Länge n bei beliebigem Kontakt zweier Teilsequenzen der Länge n einer Ribonucleinsäure
$p(kH^n)$	Wahrscheinlichkeit,dass in einer Ribonucleinsäure k verschiedene Helices der Länge n auftreten
pN	Wahrscheinlichkeit für das Auftreten der Base N an einer beliebigen Stelle einer Nucleinsäure; statt N kann A,C,G oder U stehen
pNN	Wahrscheinlichkeit für das Auftreten der Sequenz N-N in einer Nucleinsäure, entsprechend gilt für längere

	Sequenzen: pNNN,pNNNN,pNNNNNNN, etc. Statt N kann A,C,G oder U stehen
$p(N_1 \rightarrow N_2)$	Wahrscheinlichkeit, dass auf die Base N_1 die Base N_2 folgt; statt N kann A,C,G oder U stehen
pU	Wahrscheinlichkeit für das Auftreten von Uracil an einer beliebigen Stelle einer Nucleinsäure
Pr(u)	Integral der Normalverteilung von minus Unendlich bis zur Stelle u
RNS	Ribonucleinsäure
rRNS	Ribosomale Ribonucleinsäure
R17	Coliphage R17
S	Svedberg-Einheit
tRNS	Transfer-Ribonucleinsäure
t-Test	Statistischer Test, ob sich zwei Stichprobenmittelwerte signifikant unterscheiden
u	Testwert für den u-Test
U	Uracil
u-Test	Statistischer Test, ob ein Stichprobenmittelwert sich signifikant vom Mittelwert einer Grundmenge unterscheidet
X	RNS oder RNS-Abschnitt
$\bar{x}$	Mittelwert einer Stichprobe
ZW	Zwischenstück, ungebundener Abschnitt einer Ribonucleinsäure

Griechisches Alphabeth

ΔG	Gibbs freie Energie
Σ	Summenzeichen
μ	Mittelwert einer Grundmenge, aus der Stichproben entnommen werden
Π	Produktzeichen
σ	Standardabweichung einer Stichprobe
φX174	Coliphage φX174

Sonstige Zeichen

4.73 In dieser Arbeit wird in Angleichung an Computerausdrucke der Punkt als decimales Komma verwendet

== kompatibel, z.B. $(N \cdot N)_i == (N \cdot N)_j$ Basenpaar Nummer i ist kompatibel zu Basenpaar Nummer j

=/= inkompatibel, z.B. $(N \cdot N)_i =/= (N \cdot N)_j$ Basenpaar Nummer i ist inkompatibel zu Basenpaar Nummer j

-A-G-
-U-C-

oder

} Teil einer Doppelhelix

$\emptyset$ leere Menge

$\neg$ logisches NICHT

$\wedge$ logisches UND

$\vee$ logisches ODER

1 Zusammenfassung

In den letzten Jahren ist es erstmalig gelungen längerkettige RNS und DNS-Moleküle zu sequenzieren. Angeregt durch diese Entwicklungen wird in dieser Arbeit versucht, die Potenzen von einsträngigen RNS-Molekülen mit bekannter Sequenz zur Bildung von stabilen Sekundärstrukturen mathematisch zu erfassen. Nach einer Einführung in die neueren Fortschritte der RNS-und DNS-Sequenzierung werden die zwei in ihrer Nucleotidsequenz vollständig bekannten Bakteriophagen MS2 und ϕX174 näher beschrieben. Sodann wird auf die experimentellen Untersuchungen der Sekundärstruktur von RNS-Molekülen eingegangen. Hierbei entsteht die Frage, ob es möglich ist, die Sekundärstruktur einer RNS allein bei Kenntnis ihrer Sequenz und unter Berücksichtigung thermodynamischer Faktoren vorherzuberechnen. Auch soll untersucht werden, ob bekannten NS-Sequenzen eine Anpasssung an eine, eventuell vorhandene, Notwendigkeit zur Bildung stabiler Sekundärstrukturen aufweisen. Zur Beantwortung dieser Fragen werden gewisse quantitative Eigenschaften, wie z.B. die Basenzusammensetzung und nearest-neighbour-Analysen herangezogen. Hieraus ergeben sich die ersten Hinweise darauf, dass bei den untersuchten Bakteriophagen höhere Faltungskapazitäten auftreten als dies bei einer zufälligen Basenfolge zu erwarten wäre. Dies bestätigt sich auch bei der statistischen Auswertung der Auszählung aller möglichen antiparallelen Sequenzabschnitte der Phagen. Die Funktion der RNS, als genetisches Material zu dienen, wird dadurch nicht beeinträchtigt.
Es wurden weiterhin zwei Optimierverfahren zur Ermittlung der stabilsten Sekundärstruktur einer RNS bei bekannter Nucleotidsequenz entwickelt. Das erste beruht auf einer einfachen Optimumwahl aus der Menge der antiparallelen Abschnitte und das zweite permutiert diese Abschnitte, so dass alle möglichen Sekundärstruk-

turen berechnet werden und man hieraus die Stabilste auswählen kann. Da es nicht rationell ist, sämtliche möglichen Sekundärstrukturen zu berechnen, wurden in den Algorithmus verschiedene Abfragen zur Eliminierung von nachweislich nicht optimalen Strukturen eingebaut. Im Laufe der Untersuchung zeigt sich, dass die hypothetischen Sekundärstrukturen stark von der Länge des betrachteten Sequenzabschnittes abhängig sind. Ausserdem ergibt sich meist keine eindeutige optimale Struktur, sondern eine Reihe verschiedener Strukturen, die als die stabilsten anzusehen sind, und thermodynamisch nicht unterscheidbar sind.

2 Einleitung

Im Laufe der letzten Jahre ist es zunehmend möglich geworden, die Primärsequenzen von DNS und RNS-Molekülen zu bestimmen. Im Jahre 1965 analysierten HOLLEY et al. (17) die Nucleotidsequenz der Hefe-Alanin-tRNS. Seitdem wurde eine ganze Reihe weiterer tRNS-Sequenzen aufgeklärt, so dass es sogar möglich wurde, die Primärsequenzen homologer Moleküle aus verschiedenen Organismen sowie die Primärsequenzen der einzelnen tRNAs eines Organismus untereinander zu vergleichen.
Erst ab 1975-1976 ist es möglich, auch längerkettige RNS-Moleküle zu sequenzieren. Dies geschieht vorwiegend durch Anwendung kontrollierter enzymatischer Spaltung des Moleküles.
Schon bei der Untersuchung der tRNS-Moleküle ergab sich, dass eine RNS ähnlich wie ein Protein in vivo eine reproduzierbare Sekundär- und Tertiärstruktur besitzt, also nicht etwa einfach linear oder irregulär geformt ist.
Eine Vorstellung von einer möglichen Konformation eines einsträngigen Nucleinsäuremoleküles vermittelt Abbildung 1:

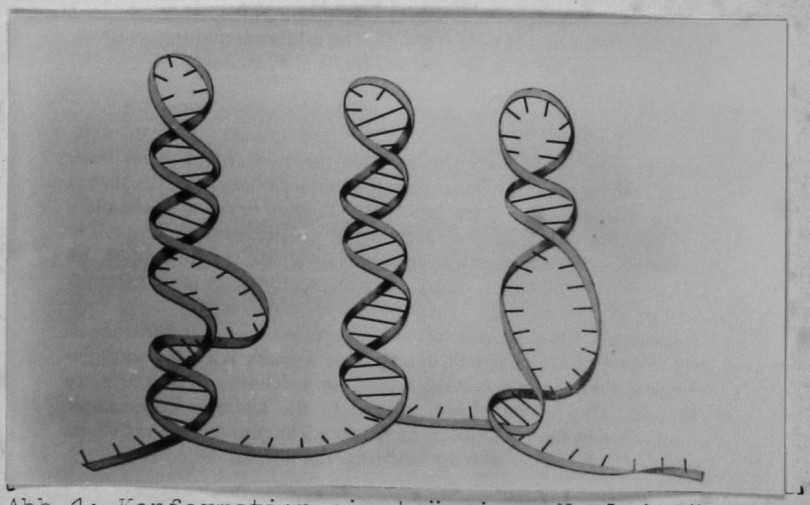


Abb.1: Konformation einsträngiger Nucleinsäuren. Entnommen aus(25a) S.13.

Durch Zusammenlagerung komplementärer Teilabschnitte des Moleküles entstehen helicale Bereiche. Dies kann sowohl bei DNS wie auch bei RNS geschehen. In der Primärsequenz der NS sind solche Konformationen durch das Auftreten hintereinanderliegender komplementärer antiparalleler Abschnitte gekennzeichnet. Ein oft erhaltenes Modell für tRNS-Moleküle ist das sogenannte "clover-leaf" (Kleeblatt) Modell (siehe Abb. 2), das sich in einzelnen Fällen auch experimentell durch Röntgenstrukturanalyse verifizieren liess (18,23). Je länger ein sequenziertes Molekül ist, desto mehr antiparallele komplementäre Abschnitte besitzt es, und desto schwieriger wird es eine eindeutige oder optimale Sekundärstruktur aufzustellen, da sich die meisten helicalen Bereiche mit anderen überlappen und es schwierig wird zu entscheiden, welche vorrangig gebildet werden und alle anderen ausschliessen.

2.1 Historische Bemerkungen zur Sequenzierung von Nucleinsäuren

Im Jahre 1965 veröffentlichten HOLLEY et al. (17) die erste vollständige Primärsequenz eines RNS-Moleküles, der transfer-RNS für Alanin aus Hefe (siehe Abb.2). Diese Sequenz wurde analysiert durch Reingewinnung des Moleküles, seiner Fragmentierung und nachfolgender Auftrennung der Bruchstücke durch DEAE-Zellulose-Säulenchromatographie.

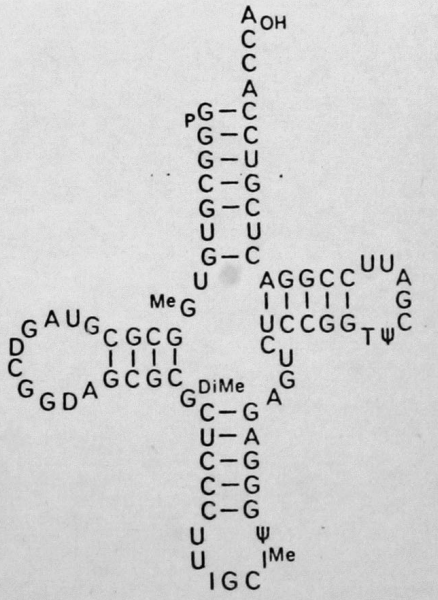


Abb. 2: Nucleotidsequenz Hefe-Alanin-tRNS. I = Inosin, Ψ =Pseudouridin, D = Dihydrouridin, I^{Me} = 1-methyl-Inosin, ^{Me}G = 1-methylguanosin, G^{DiMe} = N_2N_2-dimethylguanosin. Entnommen aus (36a).

Einen methodischen Fortschritt brachte die radioaktive Markierung der RNS-Moleküle und die Auftrennung enzymatischer Bruchstücke durch 2-dimensionale Chromatographie auf modifiziertem Papier, die durch SANGER und BROWNLEE eingeführt wurde (siehe (5)). Dieser Arbeitsgruppe

gelang es 1967 als bis dato längste RNS-Sequenz die 120 Nucleotide umfassende 5S ribosomale RNS von E.coli zu sequenzieren. Weitere nicht-tRNA-Sequenzen werden bekannt, in Abb. 3 ist z.B. die Sequenz und hypothetische Sekundärstruktur der 4.5S RNA_I aus Hepatomzellen wiedergegeben. Als geeignete Ogjekte

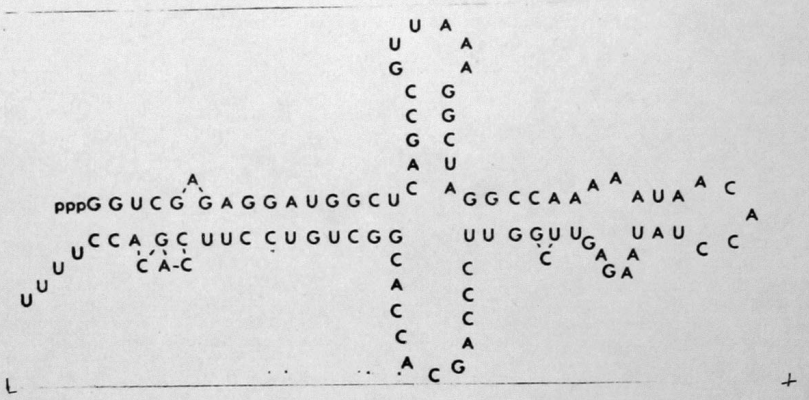


Abb. 3: Nucleotidsequenz und mögliche Sekundärstruktur der 4.5S RNA_I aus Novikoff Hepatomzellen (entnommen aus (36a) S.48).

zur Aufklärung längerkettiger RNS-Sequenzen erwiesen sich die RNS-Genome von einigen Bakterio-phagen, die sich mit hoher Ausbeute radioaktiv markieren lassen und relativ leicht rein darzustellen sind. 1969 veröffentlichte die Arbeitsgruppe SANGER (1) ein 57 Nucleotide langes Teilstück aus dem Genom des Bakteriophagen R17 (Hüllproteinregion). Die Arbeitsgruppe FIERS konnte 1972 die RNS-Sequenz des Hüllproteingenomes des Bakteriophagen MS2 vorstellen (31) und 1975 das Genom des A-Proteines (11).Im April 1976 war die Sequenzierung aller drei Genregionen (A-Protein,Hüllprotein und Replikase) einschliesslich der zwischen den Genen liegenden offensichtlich nicht mitübersetzten Zwischenstücke (abgeschlossen). Die gesamte RNS des Bakteriophagen MS2 ist 3569 Nucleotide lang (12).

Die Sequenzierung von DNS-Molekülen war anfangs weniger zugänglich als bei RNS-Molekülen, weil die Zelle keine grösseren Fraktionen kurz-

kettiger, homogener DNS-Moleküle aufweist, die in ihrer Länge etwa den transfer-RNS-Molekülen entsprechen würden. Die ersten Sequenzierungen wurden am DNS-Bakteriophagen φX174 vorgenommen. Es wurde die Depurinierungsmethode von BURTON und PETERSON (6,7) verwendet, mit der 1960 Sequenzen bis zu einer Länge von 10 Nucleotiden aufgeklärt werden konnten. In den Jahren 1973-1974 gelang die Charakterisierung längerer Teilsequenzen durch partiellen Abbau der DNS durch die Endonuclease IV und andere Enzyme. Die längste so erhaltene Sequenz besass 48 Nucleotide (14,41). Sodann wurde die Methode der Starter-abhängigen Synthese von komplementären DNS-Molekülen mit Hilfe der DNS-Polymerase entwickelt (1973-1974) (34,35), aus der die heute gebräuchliche "Plus und Minus"-Methode hervorging. Mit letzterer Methode lassen sich, ausgehend von einer kurzen Startersequenz längere Genabschnitte schnell und effektiv sequenzieren. November 1976 konnten BARREL et al. (3) die Sequenzen der Gene D und E des Bakteriophagen φX174 vorstellen und im Februar 1977 die komplette, 5375 Nucleotide lange Sequenz von φX174(33). Im Laufe des Jahres 1977 wurden weitere DNS-Sequenzen bekannt, wie z.B. von einigen Genen des simian virus 40 (SV40) und einzelner Genabschnitte aus Bakterien (20,36).

Aus den DNS-Sequenzen lassen sich bei Kenntnis der entsprechenden Transcriptionsinitiator- und -terminatorregionen leicht die Sequenzen der zugehörigen mRNS-Moleküle bestimmen.

2.2 Historische Bemerkungen zur Sekundärstrukturaufklärung

Die Sekundärstruktur doppelsträngiger Nucleinsäuren stellt eine einfache rechtsgewundenen Doppelhelix dar. Diese bildet die strukturelle Grundlage für die Fähigkeit der biologischen Organismen zur identischen Selbstreplikation und sichert deren genetische Stabilität. Die Sekundärstruktur von einsträngigen Nucleinsäuren ist weniger einheitlich. Sollte man annehmen, dass sich derartige Moleküle irregulär zusammenfalten oder haben sie eine reproduzierbare Sekundärstrukturbildung, wie etwa Proteine?
Schon bald nach Bekanntwerden der ersten Primärsequenzen von transfer-RNS-Molekülen zeigte sich, dass diese Moleküle intramolekulare, komplementäre Sequenzabschnitte besitzen, die sich bei Aneinanderlagerung haarnadelähnlich zu Doppelhelices umformen könnten. Man stellte für tRNS-Moleküle ein "Kleeblatt"-Modell ihrer Sekundärstruktur auf, dass aus mehreren haarnadelförmigen Abschnitten bestand (siehe Abb. 2 und 3). Schon bald konnte dieses Modell für einige tRNS-Moleküle experimentell durch Röntgenstrukturanalyse verifiziert werden (23,18) und es ergaben sich 3-dimensionale Modelle ähnlich Abb. 4. Auch zeigte sich, dass ribosomale RNS eine reproduzierbare Sekundärstruktur aufweisen muss, da es gelingt, ein desintegriertes Ribosom nach Denaturierung seiner Bestandteile wieder zu einem funktionsfähigen Komplex zusammenzusetzen, der dieselben Eigenschaften aufweist wie der native Komplex. (G·C)-Basenpaare sind thermodynamisch stabiler als (A·U)-Paare. Da die ribosomale RNS nun einen erhöhten Gehalt an Guanin und Cytosin besitzt, kann man annehmen, dass sich bildende Sekundärstrukturen stabiler sind, als bei vergleichbaren RNS-Molekülen mit niedrigeren Gehalten an Guanin und Cytosin. In letzter Zeit ist es allerdings gelungen auch bei einem Molekül mit mittlerem G und C Ge-

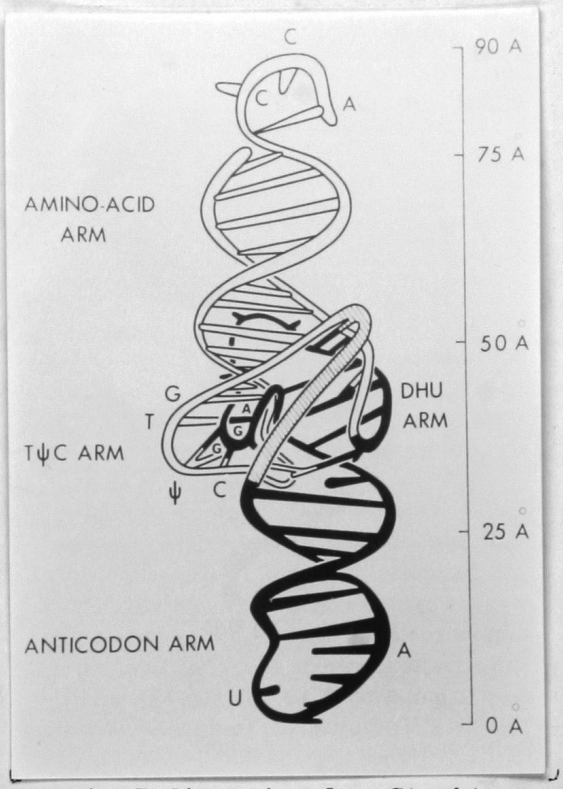


Abb. 4: 3-dimensionales Strukturmodell für tRNS-Moleküle.
Entnommen aus (36a) S.91.

halt (dem Bakteriophagen MS2) elektronenmikroskopisch Sekundärstruktur nachzuweisen (19). MS2 zeigt, abhängig von der Ionenstärke des Mediums, ein verschiedenes Ausmass an regelmässig auftretenden haarnadelförmigen Faltungen, sowie Ausbildung von offenen Schleifen. Bei der Sequenzierung von MS2 ist man desweiteren darauf aufmerksam geworden, dass Nucleasen an bestimmten Stellen des Moleküles bevorzugt angreifen. Diese Stellen müssen sich durch ihre Sekundärstruktur von anderen Stellen mit gleicher oder ähnlicher Sequenz unterscheiden.

2.3 Bedeutung der RNS-Sekundärstruktur

Die Bedeutung der RNS-Sekundär- und Tertiärstruktur liegt vor allem darin, dass nur unter ihrer Aufrechterhaltung die Funktion des Moleküls gesichert ist. Bestimmte Teile der Konformation von transfer-RNS-Molekülen werden von den jeweiligen transfer-RNS-Aminosäuren--Ligase Enzymen erkannt; die tRNS müssen der Form ihrer Bindungsorte an den Ribosomen entsprechen usw.. Pro-tRNS-Moleküle werden durch Endonucleasen in die endgültigen tRNS-Moleküle gespalten. Auch dies muss durch eine spezifische Sekundärstruktur gesteuert sein, damit keine Spaltungen an falschen Stellen geschehen. In Abb. 5 ist eine hypothetische Sekundärstruktur für die Pro-Tyrosin-tRNS aus E. coli wiedergeben, bei der eine Spaltung in der Nähe von Position Nummer 90 erfolgen muss. Auch bei

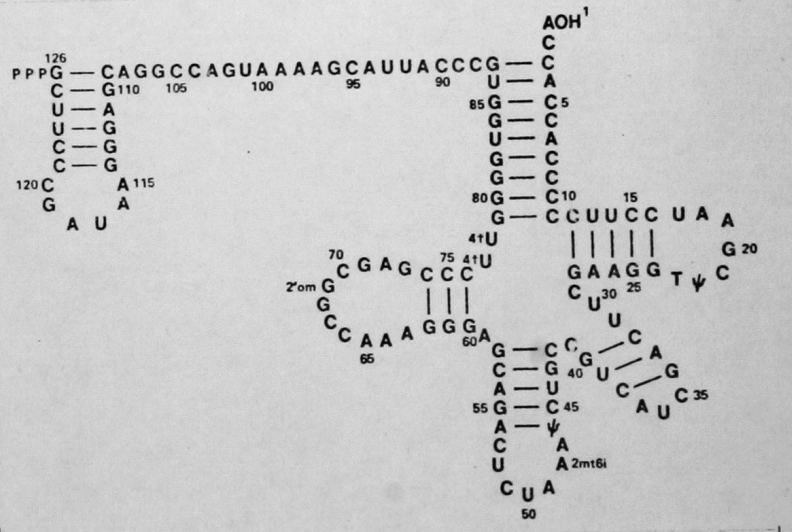


Abb.5:Struktur und Sequenz des Vorläufers der Tyrosin-tRNS aus E.coli. Abkürzungen der modifizierten Basen wie in Abb.2 Ausserdem: 2mt6iA = N^6-isopentenyl-2-methylthioadenosin, 2'omG = 2'-methoxyguanosin, 4tU = 4-thioUridin . Die modifizierten Basen finden sich nur in den funktionsfähigen tRNS-Molekülen, bei denen die Abspaltung des Restes erfolgt ist.

messenger-RNS-Molekülen hat die Sekundärstruktur mehrere Funktionen: Es ist anzunehmen, dass Endonucleasen bestimmte Bereiche in Promessengermolekülen entdecken und spezifisch spalten, bevor der an der DNS abgelesene Messenger aus dem Kern der Euzyte in den Zytoplasmabereich eintritt und dort weiterverarbeitet wird. Auch im Zytoplasma wird der Messenger weiter abgebaut. Man kann quasi von einer biologischen Halbwertszeit dieses Moleküles sprechen. Die Geschwindigkeit des Abbaus ist wiederum bestimmt von der Sekundär-und Tertiärstruktur. Die Stellen eines Messengers, die für den Start eines Translationsvorganges bestimmt sind, sind nicht nur vom Vorhandensein des Initiationscodons A-U-G abhängig, denn ein Messenger, wie z.B. MS2 besitzt bedeutend mehr A-U-G-Codons als Translationsinitiationsstellen, so dass auch hier angenommen wird, dass es zusätzlich bestimmte räumliche Strukturen des Moleküls sind, die vom Ribosom und den Initiationsfaktoren erkannt werden. Es könnte z.B. sein, dass ein A-U-G-Codon in der ungebundenen Schleife eines haarnadelförmigen helicalen Bereiches liegt und somit besonders hervorgehoben wird. Die Rate, mit der die Ribosomen einen Messenger ablesen können, kängt vermutlich ebenfalls von der Sekundärstruktur ab, denn es lässt sich leicht vorstellen, dass eine intensive und stabile Faltung des Moleküls zu helicalen Bereichen einen Ablesevorgang, der einsträngig vorliegende RNS benötigt, behindert (16). Nur durch gelegendliche thermische Auffaltungen kann die Translation ermöglicht werden und diese sind umso seltener, je stabiler eine bestimmte Konformation ist. Nicht alle Proteine, die auf einem Messenger codiert sind, werden in gleicher Menge und zu gleichen Zeitpunkten abgelesen. Auch diese Vorgänge können durch die Konformation des Moleküls gesteuert sein.

2.4 Übersicht über bisherige Verfahren zum Auffinden hypothetischer Sekundärstrukturen von RNS

Bei dem Versuch zu hypothetischen Sekundärstrukturen zu gelangen, die zumindest angenähert denen entsprechen, die in vivo vorliegen, git es zwei grundsätzlich verschiedene Ansätze, je nach dem, ob man von der bekannten Primärsequenz eines RNS-Moleküles ausgeht, oder ob man versucht über die bekannte Sequenz eines Proteines auf die Sequenz des entsprechenden mRNS-Moleküles zu schliessen. Mit dem ersten Problem befassten sich im Jahre 1971 TINOCO, UHLENBECK und LEVINE (38). Sie stellten fest, welche Basen miteinander eine Bindung eingehen und sich am Aufbau einer Doppelhelix beteiligen könnten. Es zeigte sich, dass zusätzlich zu den klassischen WATSON-CRICK-Basenpaaren (G·C) und (A·U) auch G und U eine solche Bindung eingehen können. Dieses wird als Wobble-Bindung bezeichnet und hat eine thermodynamisch geringere Stabilität als die beiden anderen Basenpaarformen. (Es ist übrigens in dieser Arbeit, wenn von "Bindung","Kopplung","Zusammenlagerung" u.ä. gesprochen wird immer die Basenpaarbindung gemeint. Da sich bei der Veränderung und Bildung von Konformationen eines Moleküls keine echten chemischen Bindungen verändern, kann hiermit keine Verwechslung auftreten.) Fernerhin stellten sie mit dieser Kenntnis die Matrix aller möglichen Basenpaarbildungen für eine bestimmte RNS-Sequenz auf, aus der sie dann "von Hand" oder "by inspection" versuchten eine möglichst stabile Sekundärstruktur abzuleiten. Im Jahre 1974 werteten GRALLA und DELISI mit derselben Methode erhaltene Sekundärstrukturen von Zufallspolymeren aus (15). Die Zufallspolymere hatten für jede Base eine gleich grosse mittlere Wahrscheinlichkeit. Es zeigte sich, dass auch bei solchen Molekülen, die keinerlei Restriktionen durch eine eventuell zu codie-

rende Aminosäurensequenz unterliegen, ausreichend stabile Sekundärstrukturen möglich waren, deren Basen durchschnittlich zu über 50% an der Basenpaarbildung beteiligt waren. Die untersuchten Sequenzen hatten eine Länge von 77 Nucleotiden, was der durchschnittlichen Länge einer transfer-RNS entspricht.

Es wurden auch verschiedene Versuche gemacht den Einfluss der subjektiven Auswahl von Sekundärstrukturen zu vermindern, indem man dazu überging, solche Strukturen vom Computer maschinell erstellen zu lassen. 1971 Stellte JORDAN ein Programm auf (21), dass nach einer "Monte-Carlo"-Methode aus der Menge aller möglichen Sekundärstrukturen zufällig welche herausnahm. Leider produzierte das Programm auch Strukturen, die zwar ein genügendes Ausmass an Basenpaarbindung beinhalteten, die aber sicher aus sterischen Gründen als unmöglich anzusehen sind. Die längste untersuchte RNS war das 120 Nucleotide lange 5S ribosomale RNS-Molekül.

Kürzlich (1977) wurde von LAPIDUS und ROSEN (26) über die Aufstellung eines Programmes berichtet, dass in der Lage ist, für ein Molekül mit vorgegebener Sequenz diejenige Sekundärstruktur aufzufinden, die nachweislich das grösste Ausmass an Basenpaarbindung besitzt. Es wurden Teile des Genoms des Bakteriophagen Qβ untersucht. Soweit aus der Arbeit hervorgeht besitzen die bearbeiteten Sequenzabschnitte eine Maximallänge von etwa 200 Nucleotiden. Da das Programm allein auf maximale Basenpaarbindung hin optimiert und keine Rücksicht nimmt auf die thermodynamischen Stabilitäten der gebildeten Sekundärstrukturvorschläge, ist wahrscheinlich in dieser Hinsicht kein Optimum erreicht worden.

1975 wurden von McMAHON und PIPAS drei Programme entwickelt, die zur Untersuchung von transfer-RNS-Molekülen herangezogen wurden. Das

erste Programm stellte alle möglichen helicalen Bereiche fest, die sich bei einer gegebenen Sequenz bilden könnten. Das zweite Programm versucht dann sämtliche sterisch möglichen Kombinationen dieser helicalen Bereiche zu finden und das dritte Programm berechnet sodann die thermodynamischen Stabilitäten der generierten Strukturen (29,32). Als mögliche helicale Bereiche wurden nur solche berücksichtigt, die mindestens eine Länge von drei Basenpaaren hatten. Auf diese Weise wurde versucht, diejenige Struktur zu finden, die bei den aus der Literatur bekannten Werten über die thermodynamische Stabilität den maximalen Gesamtstabilitätswert besitzt, bzw. die Klasse der Strukturen, die dem Maximum am nächsten kommt. Wieder wurden nur RNS-Moleküle von der Länge der transfer-RNS untersucht. Faltungen können auch helicale Bereiche mit weniger als drei Basenpaaren enthalten. Hieraus und aus der Beschreibung der Programme geht hervor, dass manche möglichen Sekundärstrukturen nicht aufgefunden werden können. Die Autoren versuchen seitdem ihre Programme auch für längerkettige RNS-Bearbeitungen zu erweitern.

Beim Versuch, von der Sequenz eines Proteins auf die Sequenz und Sekundärstruktur eines entsprechenden Messengers zu schliessen, besteht die Unsicherheit in der Wahl der dritten Base eines Codons für eine Aminosäure, da die meisten Aminosäuren von mehreren Tripletts codiert werden, die sich in den dritten Basen unterscheiden. Dies nennt man auch die Degeneration des genetischen Codes und die dritte Base heisst die degenerierte dritte Position. Die dritte Position kann nun an der Bildung bestimter Sekundärstrukturen beteiligt sein.Sie kann so gewählt sein, dass im Gesamtmolekül entweder möglichst wenig Basenpaarbildung auftritt oder möglichst viel.

FIGUEROA,SOTO et al.(1972) (13) nahmen an, dass es für einen Messenger am vorteilhaftesten ist,wenn er eine möglichst geringe Sekundärstruktur besitzt, da dadurch der Translationsvorgang erleichtert würde. Sie machten aufgrund dieser Annahme eine Vorhersage über die mögliche Sequenz des Messengers für Cytochrom C aus dem menschlichen Herzen.
LAUX,DENIS,WHITE (27) sowie KLÄMBT und RICHTER(24, 25) gehen davon aus, dass eine gut ausgebildete Sekundärstruktur den Messenger vor Abbau durch Nucleasen schützt und gewisse Funktionen beim Erkennen der Initiations- und Terminationsstellen erfüllt. Unter Verwendung dieser Annahme versucht~~en~~ Sie zu einer Vorhersage für die Primärsequenz des Messengers für die α-Kette des menschlichen Globins zu kommen.
Die Methodik der letzten Arbeitsgruppen ist sehr unterschiedlich. FIGUEROA,SOTO et al. vergleichen die Häufigkeiten von möglichen helicalen Bereichen mit einer Mindestlänge von vier Nucleotiden bei verschiedenen denkbaren Messenger-Molekülen. LAUX,DENIS und WHITE versuchen, ausgehend von einem bekannten und sequenzierten Teil des Messengers auf einen unbekannten zu schliessen, von dem sie annahmen, dass er sich mit dem bekannten zu einer haarnadelförmigen Struktur zusammenlagern würde. KLÄMBT und RICHTER liessen ein Computerprogramm ~~su~~k essive nebeneinanderliegende "Haarnadeln" generieren. Leider war es in letzterem Programm nicht möglich, Bindungen zwischen weiter auseinanderliegenden Teilen des Moleküles herzustellen und die einzelnen Haarnadeln enthielten keine Bindungsfehler oder "mismatchings", die zu einem Herausklappen von einzelnen Basen aus einer fortlaufenden Helix führen. Somit wurden nicht alle möglichen Strukturen oder "Faltungen" generiert.

2.5 Beschreibung des Bakteriophagen MS2

MS2 besteht aus einer einsträngigen RNS-Kette, die in gefalteter Form in einem isometrischen Virion verpackt ist, dass aus dem Reifungsprotein und etwa 180 Einheiten des Hüllproteines besteht. Es befällt das Enterobakterium Escherichia coli. Sein RNS-Molekül enthält 3569 Nucleotide und codiert drei virus-eigene Proteine, das Reifungsprotein, das Hüllprotein und das Replikaseprotein, in der genannten Reihenfolge. Seine Primärsequenz wurde vollständig aufgedeckt und ist dem Anhang zu entnehmen. Siehe auch (31,11,12). Das Reifungsprotein, auch Leitprotein (pilot protein), A-Protein oder Reifungsfaktor (maturation factor) genannt, hat eine Länge von 393 Aminosäuren, eine Molekularmasse von ungefähr 43 KDalton und wird von Nucleotid 130 bis Nucleotid Nummer 1308 codiert. Es ist essentiell für die Bildung eines infektiösen Phagen und liegt in einer Kopie pro Phage vor. Ohne Reifungsprotein sind die Phagen-RNS-Moleküle anfällig gegenüber dem Abbau durch die RNase I und elektronenoptische Bilder zeigen ein Heraus"schlenkern" der RNS aus dem Phagenpartikel. Es ist anzunehmen, dass das in wässriger Phase schwer lösliche Protein bei der Anheftung des Phagen an die Wirtzelle behilflich ist und möglicherweise eine Pore für das Eindringen der RNS in die Zelle bildet. Auch wird vermutet, dass das Protein als Leitprotein in der Lage ist, die Spezifität der wirtseigenen Translationsmaschinerie zu Gunsten der Phagen-RNS zu verändern. Das Hüllprotein (coat protein) hat eine Länge von 129 Aminosäuren, eine Molekularmasse von ungefähr 14 KDalton und wird von Nucleotid 1335 bis Nucleotid 1724 codiert. Es ist ebenfalls schwerlöslich in Wasser und liegt im reifen Phagen in etwa 180 Kopien vor und stellt somit den Hauptteil der Protein-

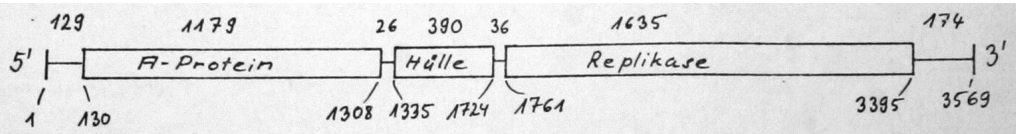


Abb. 6:Das Genom von MS2 enthält drei Gene, die als Rechtecke dargestellt sind. An beiden Enden und zwischen den Genen befinden sich nicht übersetzte Abschnitte.Oben sind die Längen der einzelnen Abschnitte ausgedrückt durch die Anzahlen der Nucleotide eingetragen. Die Nucleotide sind von 5'bis zum 3'-Ende durchnummeriert und die Positionen einiger wichtiger Stellen sind unten wiedergegeben. (Entnommen aus (12)).

hülle des Phagen dar. Die Replikase, auch Virus-RNS--abhängige-RNS-Polymerase oder RNS-Synthetase genannt, hat eine Länge von 544 Aminosäuren, eine Molekularmasse von ungefähr 59 KDalten und wird von Nucleotid 1724 bis Nucleotid 3395 codiert. Der Replikase-Komplex, der für die Vermehrung des Phagen in der Zelle zu sorgen hat, besteht allerdings aus vier Komponenten, unter denen die vom Phagen codierte "Replikase" nur eine ist. Die Komponenten werden α, β, γ und δ genannt. α erwies sich als Bestandteil S1 der 30S ribosomalen Untereinheit, β stellt die "Replikase" selbst dar und γ und δ konnten als die Elongationsfaktoren Tu und Ts identifiziert werden. Die Abb. 6 vermittelt eine Vorstellung von der Lage der drei Gene auf der PhagenRNS.

Die Translation der drei Proteine des Phagen erfolgt nicht gleichzeitig und in gleicher Menge, sondern ist gesteuert. Das Reifungsprotein wird vermutlich nur von nascenten noch nicht voll replizierten Phagen abgelesen und die Translation der Replikase ist abhängig von der Translation des Hüllproteins; erst wenn dieses abgelesen worden ist,kann die Translation der Replikase initiiert werden. Ferner unterdrückt das fertige Hüllprotein die Translation der Replikase, indem es sich mit der RNS im Bereich des Replikase-Genes zusammenlagert.

2.6 Beschreibung des Bakteriophagen φX174

φX174 ist ein Bakteriophage, der die Enterobakterien Escherichia coli und Salmonella typhimurium befällt. Es handelt sich um ein ikosaedrisches Polyedron, dessen 12 Ecken mit Nadeln oder "Spikes" besetzt sind (Siehe Abb. 7). Sein Genom besteht im Gegensatz zu MS2 nicht aus RNS, sondern aus DNS mit einer Länge von 5375 Nucleotiden, auf der sich die Gene für neun virus-

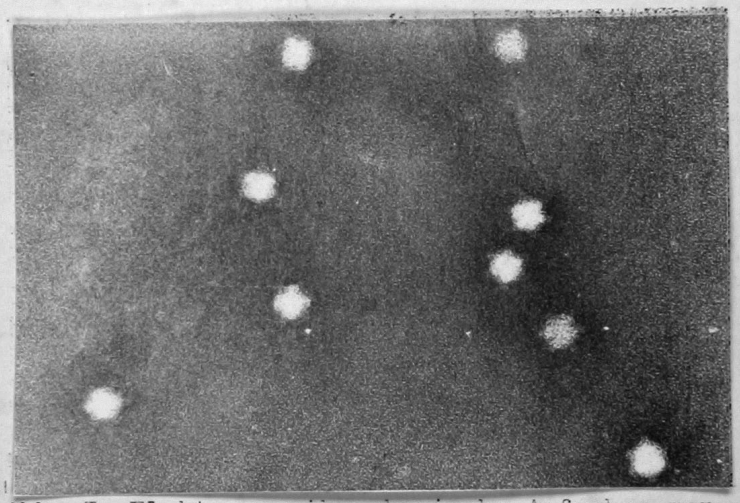


Abb. 7: Elektronenmikroskopische Aufnahme von X174. Vergrösserung 200000fach. Entnommen aus(25a).

eigene Proteine befinden, die mit den Buchstaben A bis I bezeichnet werden:

	Mol.gew. nach SDS-Meth.	Anzahl codierender Nucleotide	Mol.gew. nach Sequenz
A	55-67000	1536	56000
B	19-25000	360	13845
C	7000		
D	14500	456	16811
E	10-17500	273	9940
J	5000	114	4097
F	48000	1275	46400
G	19000	525	19053
H	37000	984	35800
Nicht-codierende Teile + C		485	

(Zusammengestellt nach (33))

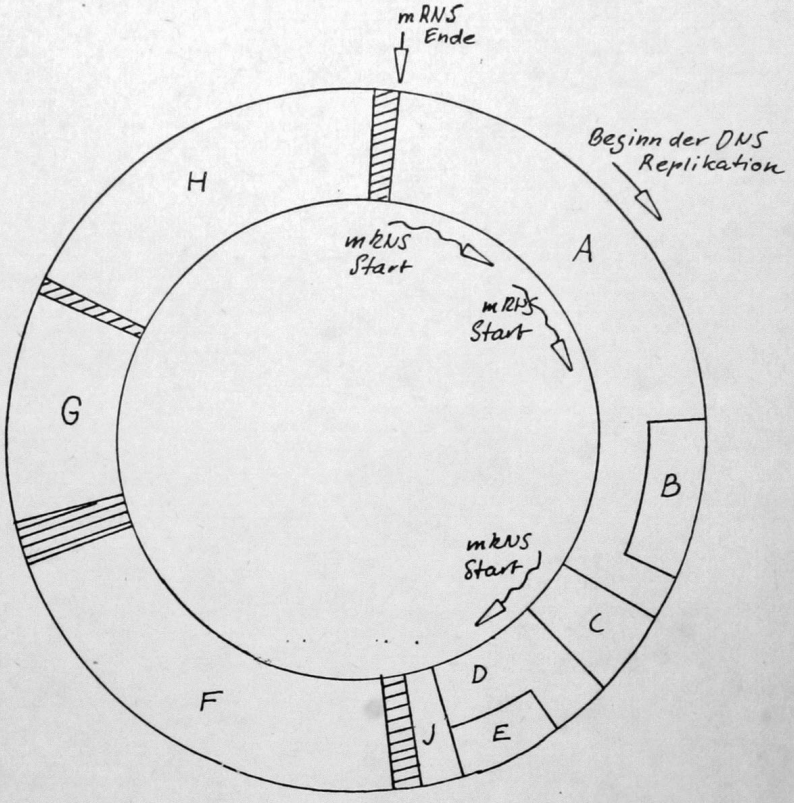


Abb.8: Genkarte des Bakteriophagen φX174. Nicht-codierende Abschnitte wurden schraffiert. Zusammenstellung nach (33).

Sie liegen auf einem ringförmigen Genom, wie dies der Zeichnung (Abb. 8) zu entnehmen ist. Besonders zu bemerken ist, dass sich die Gene für die Proteine A und B sowie die Gene für die Proteine D und E überlagern, es werden hier verschiedene Proteine vom selben Genort unter Verwendung verschiedener Leserahmen (reading frames) abgelesen. So sind z.B. die Tripletts für die Aminosäuren von Protein E um eins gegenüber den Tripletts für die Aminosäuren von Protein D verschoben. Man kann annehmen, dass die Sequenzen von sich überlagernden Genen stark voneinander abhängig sind. Wie stark diese Abhängigkeit allerdings ist und in welchem Ausmass dadurch die Funktionen der Proteine beeinflusst werden, steht noch nicht fest.

Die Ribosomenbindungsorte von φX174 zeichnen sich

im allgemeinen durch eine, dem eigentlichen Initiationscodon vorgelagerte, Polypurinsequenz aus, von der angenommen wird, dass sie mit dem 3'-Ende der 16S rRNS Basenpaarbindung eingehen kann. Die Proteine F,G und H stellen Strukturproteine für die Hülle des Phagen dar. Der Phage besitzt 60 F Proteine pro Virion und jeder seiner 12 Spikes besteht aus 5 G Proteinen und einem H Protein deren Anordnung man dem Schema in Abbildung 9 entnehmen kann.

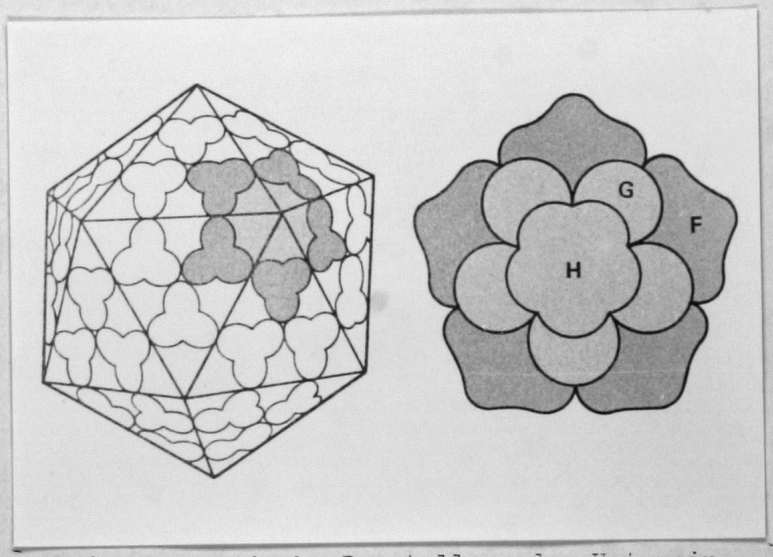


Abb. 9: Schematische Darstellung der Untereinheiten von ΦX174(links) und eine Vergrösserung der Spike-Region mit den vermutlichen Anordnungen der Proteine H,G und des Haupthüllproteines F. Entnommen aus (25a)

Das H Protein hat bei ΦX174 vermutlich die Funktion eines Leitproteines (pilot protein), da es wahrscheinlich mit der DNS in die Wirtszelle eindringt. I ist ein kleines basisches Protein, das ebenfalls Teil des Virions ist. A wird zur Doppelstrangreplikation und zur Synthese der Einzelstränge benötigt. Die Proteine D,C und D tragen in noch unbekannter Weise zur Produktion der Einzelstränge bei und helfen bei der richtigen Bündelung (packaging) der DNS. Das Protein E schliesslich führt zur Lyse der Wirtszelle.

2.7 Fragestellungen

Die vorliegende Arbeit verfolgt mehrere Ziele. Es soll zum Einen festgestellt werden, ob die bei MS2 vorhandene Kapazität zur Bildung von Sekundärstrukturen eine Besonderheit dieser RNS ist, oder ob sie bei jeder beliebigen Ribonucleinsäure, die in einsträngiger Form vorliegt in gleicher Weise erwartet werden kann. Zum Anderen muss untersucht werden, wie gross die Beschränkungen sind, die eine eventuell in vivo vorhandene Notwendigkeit zur Bildung stabiler Sekundärstrukturen, den Primärsequenzen auferlegt. Mit anderen Worten: Kann ein solcher Messenger noch in ausreichendem Masse seiner Funktion als genetischer Informationsträger gerecht werden? Ist er nicht durch die notwendigen antiparallelen komplementären Sequenzen zu sehr festgelegt, als dass er noch die, für die Bildung sinnvoller Proteine notwendige, Information liefern kann?
Zum weiteren soll versucht werden, zu einem algorithmischen Verfahren zu gelangen, das möglichst optimale Faltungen beliebiger RNS-Moleküle generiert und dabei nicht mehr von den teilweise subjektiven Kriterien bestimmt wird, mit denen die bisher in der Literatur vorgeschlagenen Sekundärstrukturen aufgefunden wurden. Eine solcherart rein "maschinell" erhaltene Faltung des MS2-Moleküles soll mit der von der Abeitsgruppe FIERS vorgeschlagenen und teilweise experimentell gestützten Sekundärstruktur verglichen werden. Hierbei wird es möglich sein, abzuschätzen, ob die numerischen Werte, die der Berechnung der Stabilität eines RNS-Moleküles zugrunde liegen, eine ausreichende Begründung für die Bildung einer bestimmten und stabilen Sekundärstruktur von RNS-Molekülen geben oder nicht. Möglicherweise ist der Einfluss weiterer Faktoren, wie z.B. die Zusammensetzung des Lösungsmittels und die Interaktionen zwischen verschiedenen helicalen Bereichen viel ausschlaggebender als bislang an-

genommen wird.
Man kann weiterhin den Algorithmus zur Erstellung von stabilen Sekundärstrukturen ausser auf MS2 auch auf hypothetische Messenger anwenden, deren Basensequenz~~en~~ mit Hilfe eines Zufallsgenerators aufzustellen wären. Hierdurch könnte man feststellen, ob und in wie weit sich bei MS2 stabilere Sekundärstrukturen auffinden lassen als bei beliebigen anderen RNS.

3 Materialien und Methoden

Die Computerprogramme wurden mit Hilfe einer Rechenmaschine des Typs IBM/370-168 am Hochschulrechenzentrum Bonn durchgeführt. Sie wurden auf Time-sharing-Basis (TSO) über eine Datenendstation des Typs IBM-2741 (Schreibmaschine) eingegeben und ausgetestet.
Die Programme wurden teils in der Programmiersprache FORTRAN IV erstellt und teils in der Programmiersprache SIMULA 67. Es zeigte sich, dass SIMULA für die Programmierung von komplexen biologischen Modellen mit relativ geringer numerischer Verarbeitung entscheidend besser geeignet ist als FORTRAN. SIMULA ermöglicht eine bessere logische Untergliederung der zu bearbeitenden Probleme.

4 Verwendete Verfahren und Ergebnisse

Die Verfahren wurden zumeist auf die Sequenz des Bakteriophagen MS2 angewendet und, soweit dies möglich war, durch analoge Anwendungen bei φX174 ergänzt.

4.1 Basenzusammensetzung und 2er Sequenzen

Weisen die bekanntgewordenen Sequenzen eine Anpassung an einen Evolutionsdruck in Richtung auf eine stabilere Sekundärstruktur auf oder nicht? Wie stark ist diese Anpassung, wenn vorhanden? Diese Fragen lassen sich z. T. allein bei Kenntnis der Basenzusammensetzung des RNS-Moleküles beantworten. Weitere Schlüsse sind möglich, wenn zusätzlich die Häufigkeiten bekannt sind, mit denen die einzelnen Basen aufeinanderfolgen. Letzteres entspricht den relativen Häufigkeiten aller möglichen 2er Sequenzen. Beides, die Information über die Basenzusammensetzung und die Kenntnis der 2er Sequenzen lässt sich auch ohne Kenntnis der genauen Sequenz gewinnen. Ersteres durch chemische Analyse und letzteres durch die sogenannte nearest-neighbour-analysis.

4.1.1 Basenzusammensetzung

Die Basenzusammensetzung der Phagen MS2 und φX174 ist wie folgt:

absolut:	U	C	A	G	Σ
MS2	875	933	834	927	3569
φX174	1677	1156	1286	1251	5370
prozentual:					
MS2	24.52	26.14	23.37	25.97	100 %
φX174	31.21	21.51	23.93	23.35	100 %

Die Wahrscheinlichkeit, mit der sich, bei Kontakt zweier beliebiger Basen des Moleküls, ein Basenpaar bildet, gibt ein ungefähres Mass für die Kapazität, stabile Sekundärstrukturen zu bilden. Diese Wahrscheinlichkeit wird in der Folge als Faltungskapazität (pF) bezeichnet.

Als Schätzwert für die Wahrscheinlichkeit des Auftretens einer bestimmten Base N (= pN) wird der Prozentsatz (geteilt durch 100) genommen, z.B. ist pA = 0.2425 bei MS2. Wenn man pF errechnen will, so muss man in Betracht ziehen, dass in der Sekundärstruktur einer RNS nicht nur die klassischen WATSON-CRICK Gasenpaare (G·C) und (A·U) auftreten können, sondern auch die Wobble-Paarung (G·U), die zwar kaum zur Stabilität eines helicalen Bereiches beiträgt, dafür aber die Symmetrie und Struktur der Helix nicht beeinträchtigt. Die Wahrscheinlichkeit für die Bildung eines bestimmten Basenpaares entspricht dem Produkt der Wahrscheinlichkeiten für das Auftreten der beiden Basen, aus denen das Basenpaar besteht. Hieraus ergibt sich für pF folgende Formel:

pF = 2pGpC + 2pApU + 2pGpU

A. M. LESK verwendet bei seiner kombinatorischen Untersuchung von RNS-Zufallspolymeren (28) die gleiche Formel. Der erste Summand von pF gibt die Wahrscheinlichkeit der Bildung eines (G·C)-Paares wieder; der Faktor 2 ergibt sich aus der Überlegung, dass entweder ein (G·C) oder ein (C·G)-Paar entstehen könnte. Beide Paare haben dieselbe Wahrscheinlichkeit,weil pGpC = pCpG ist, und man kann pGpC einfach verdoppeln.
Bei gleichförmiger Basenzusammensetzung, wenn pU = pC = pA = pG = 0.25 ist,hat pF den Wert 0.375 . Bei MS2 ergibt sich aber ein Wert von 0.37774 und bei φX174 von 0.39559 . In beiden Fällen ist also eine Steigerung der Faltungskapazität über Normalkapazität nachweisbar.

4.1.2 Struktur der Funktion pF

Um sich über den Zusammenhang zwischen pF und der Basenzusammensetzung klarzuwerden, ist es hilfreich, sie einmal graphisch darzustellen. pF hängt von vier Variablen ab, ist demnach eine Funktion im vierdimensionalen Raum. Da die Variablen pU,pC,pA und pG aber nicht völlig unabhängig voneinander sind, sondern durch die Funktion 1.0 = pU + pC + pA + pG zusammenhängen, ist es bei Einsetzen dieser Funktion möglich eine Dimension einzusparen, man erhielte somit die Möglichkeit pF im dreidimensionalen Raum wiederzugeben. Um zunächst zu einer Darstellung auf der Zeichenebene zu kommen, sei der Sonderfall pC = 0.0 untersucht. Es wird dann

pF = 2pApU + 2pGpU

Diese Funktion liegt, wie in Abb. 10 zu sehen, in einer Ebene des dreidimensionalen Raumes.

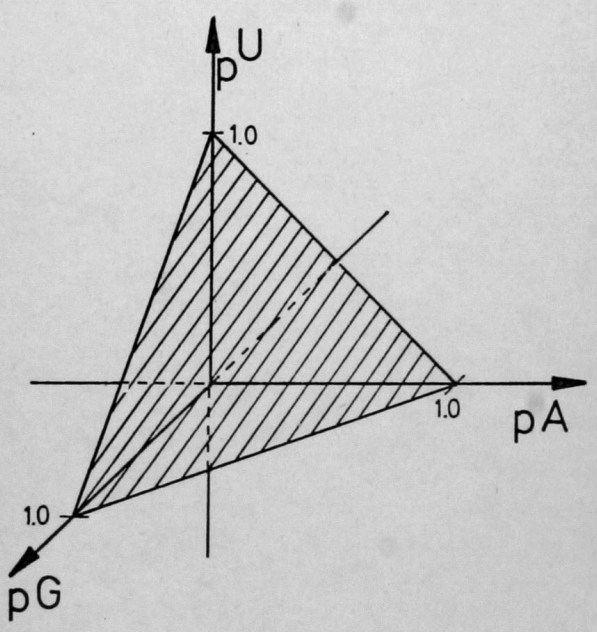


Abb.10: Graphische Darstellung der Funktion pF = 2pApU + 2pGpU

Man kann diese Ebene mit der Zeichenebene zusammenfallen lassen und erhält dann eine Darstellung wie in Abb. 11. Man sieht, dass die

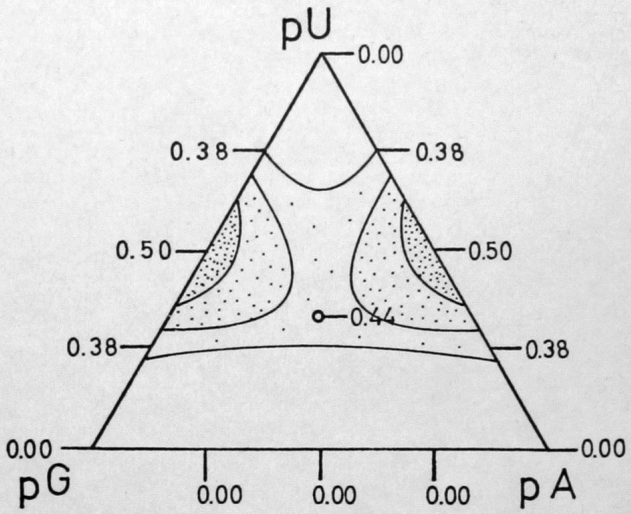


Abb. 11:Ebene Darstellung der Funktion
pF = 2pApU + 2pGpU

Maxima der Funktion auf den Kanten liegen, die Ecken gleich 0.0 werden und bei Gleichwahrscheinlichkeit der Basen der wenig unter dem Maximum 0.5 liegende Wert 0.444 erhalten wird. Eine gleiche Betrachtung lässt sich für die Fälle pA = 0.0,pU = 0.0 und pG = 0.0 durchführen. Man erhält daraufhin vier gleichseitige Dreiecke, die sich zu einem Tetraeder zusammensetzen lassen und hat somit die gesuchte dreidimensionale Darstellung von pF erhalten (Abb.12). Die Maxima der Funktion liegen auf den Kanten, die vier Ecken haben keine Faltungskapazität und der Mittelpunkt des Tetraeders, also der Fall einer Gleichwahrscheinlichkeit aller Basen hat den bereits oben erwähnten Wert 0.375. Wenn das Molekül sich also einem Evolutionsdruck in Richtung auf einen höheren Grad von Faltungskapazität beugt, wird sich sein pF--Wert in Richtung auf eine der Kantenmitten zu bewegen, auf denen die Maxima der Funktion liegen. Am wahrscheinlichsten ist ein Zuwandern auf die Mitte der G-C-Kante, weil (G·C)-Basenpaare die höchste thermodynamische Stabilität besitzen.

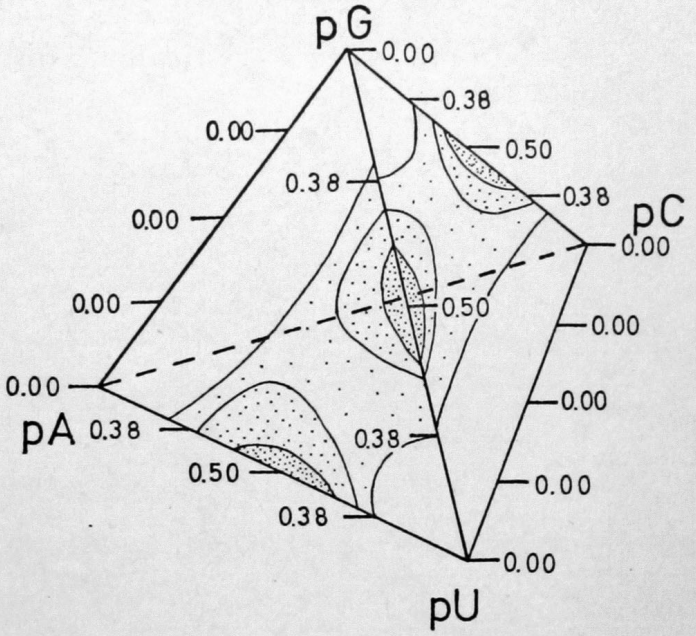


Abb. 12: Graphische Darstellung der Funktion
pF = 2pGpC + 2pApU + 2pGpU

4.1.3 Informationskapazität

Das Vermögen der RNS, als Informationsträger zu fungieren, d.h. als Speicher für Erbinformationen zu dienen, ist abhängig von seiner Basenzusammensetzung. Die Anzahl unterschiedlicher Primärsequenzen in denen die RNS auftreten kann, nimmt desto mehr ab, je mehr die Basenzusammensetzung von der Gleichverteilung abweicht. Ist z.B. $pU = pC = pA = pG = 0.25$, so lassen sich 256 verschiedene Sequenzen zu je vier Nucleotiden bilden, ist jedoch $pU = pC = pA = 0.33$ und $pG = 0.0$ so ist es nur noch möglich 81 verschiedene Sequenzen herzustellen. Von der Anzahl verschiedener möglicher Sequenzen einer Nucleinsäure ist die Anzahl unterschiedlicher Proteine bestimmt, die von ihr abgelesen werden können. Unter den nicht mehr ablesbaren Proteinen können sich auch biologisch sinnvolle befinden.

Ein quantitatives Mass für die Informationskapazität einer Nachricht - und man kann die RNS-Sequenzen auch als Nachrichten auffassen - liefert die Informationstheorie, es ist die Anzahl der bit, die pro Buchstabe oder Zeichen codiert werden können. Ist die Wahrscheinlichkeit eines Buchstabens N gleich pN, so ist nach (1a) die Informationskapazität gleich:

$$I = - \sum_i pN_i \cdot ld(pN_i)$$

wobei i über alle vorkommenden Buchstaben summiert wird und

$$\sum_i pN_i = 1$$

ist. Unter ld ist der Logarithmus zur Basis 2 zu verstehen. Im vorliegenden Fall ist

$$I = -pU \cdot ld(pU) - pC \cdot ld(pC) - pA \cdot ld(pA) - pG \cdot ld(pG)$$

Bei Gleichverteilung der Basen, wenn $pU = pC = pA = pG = 0.25$ ist, vereinfacht sich I zu:

$$I = - 4 \cdot 0.25 \cdot ld(0.25) = - ld(0.25) = + ld(4.0) = 2.0$$

Eine Base kann somit 2 bit codieren, ein Tripplet codiert 6 bit usw..

Bei MS2 liegt nach der obigen Formel der I-Wert pro Base bei 1.9985 und bei ϕX174 bei 1.9849 . Der Unterschied zwischen den errechneten I-Werten und dem Maximalwert von 2.0 ist numerisch schwer erfassbar und liegt bei MS2 unter 0.005 bit und bei ϕX174 unter 0.05 bit. Man kann also davon ausgehen, dass die Anpassung an stabile Sekundärstrukturen mit Hilfe der Veränderung der Basenzusammensetzung nur zu einer unwesentlichen Einschränkung der Codierfähigkeiten der beiden Phagen geführt hat.

4.1.4 Struktur der Funktion I

Es gelten für die Funktion I, die von den vier Variablen pU, pC, pA und pG abhängt, dieselben Überlegungen bezüglich der Dimensionalität und bildlichen Darstellbarkeit einer vierdimensionalen Funktion, wie für pF. Auch hier wird durch die Einschränkung 1.0 = pU + pC + pA + pG eine Dimension eingespart. Darum wird zunächst wieder der Sonderfall des Wegfallens einer Base betrachtet, es sei also pA = 0.0. Es ist dann

$$I = - pU \cdot ld(pU) - pC \cdot ld(pC) - pG \cdot ld(pG)$$

Die bildliche Darstellung dieses Ausdruckes entspricht der Abb.10. Die Funktion weist wiederum nur Werte in einer Ebene auf (Abb.13):

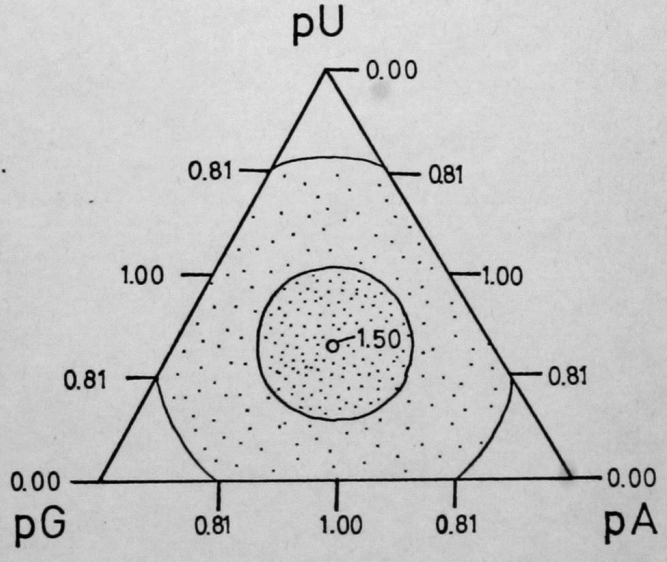


Abb.13: Ebene Darstellung der Funktion
$I = - pU \cdot ld(pU) - pC \cdot ld(pC) - pG \cdot ld(pG)$

Auch eine Zusammensetzung aller möglichen Dreieck_sflächen zu einen Tetraeder ist möglich (Abb. 14) . Es zeigt sich, dass die Funktion nur ein einziges Maximum besitzt, nämlich genau in der Mitte des Tetraeders bei der Gleichverteilung der Basen. Die Kantenmitten besitzen nur halbmaximale Werte und die Ecken weisen keine

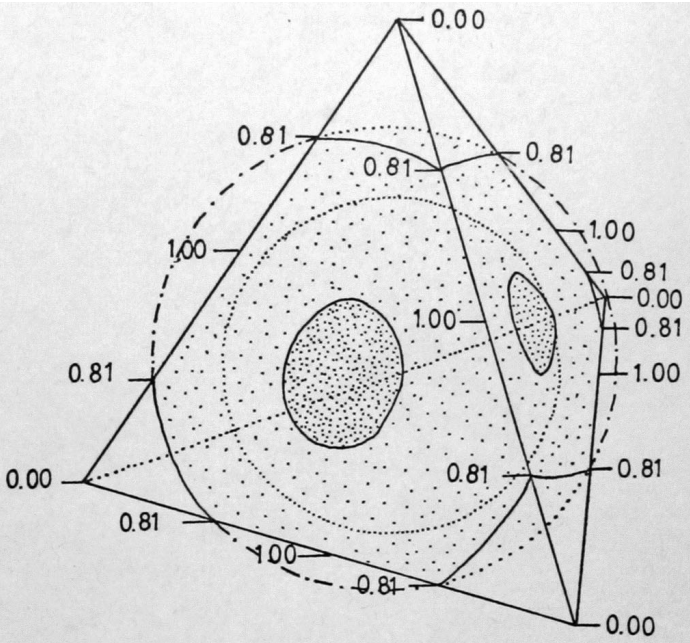


Abb. 14: Graphische Darstellung der Funktion
$I = - pU \cdot ld(pU) - pC \cdot ld(pC) - pA \cdot ld(pA) - pG \cdot ld(pG)$

Informationskapazität auf.
Die RNS unterliegt also zwei verschiedenen Evolutionsdrücken: einmal muss sie möglichst grosse Informationskapazität besitzen, strebt also zur Mitte des Tetraeders, zum anderen treibt die Notwendigkeit zur Sekundärstrukturbildung zu einer der Kantenmitten, vorzugsweise der G-C--Kante. Je nach der biologischen Funktion des betreffenden RNS-Moleküls wird es mehr nach der einen oder mehr nach anderen Seite streben.

4.1.5 2er Sequenzen

Eine RNS mit einer bestimmten Basenzusammensetzung kann eine Vielzahl von verschiedenen Sequenzen besitzen. Es ist nun möglich, dass die RNS nicht nur durch Veränderung der Basenzusammensetzung ihre Faltungskapazität erhöht hat, sondern auch durch eine Veränderung ihrer Sequenz. Die einfachste Weise dieser Frage nachzugehen, ist die Auszählung der verschiedenen möglichen 2er Sequenzen und ein nachfolgender Test, ob die Häufigkeiten der einzelnen Sequenzen gegenüber den Erwartungswerten signifikant verschieden sind und ob diese Unterschiede im Sinne einer vermehrten oder verminderten Fähigkeit zur Bildung von Sekundärstrukturen zu deuten sind.
Es sei zunächst eine Liste der 16 möglichen 2er Sequenzen gegeben, sowie eine Auszählung ihres Auftretens in MS2 und φX174:

	MS2	φX174
U-U	211(214)	568(523)
U-C	263(229)	321(361)
U-A	194(204)	312(402)
U-G	207(227)	476(391)
C-U	242(229)	403(361)
C-C	223(244)	227(250)
C-A	201(218)	259(277)
C-G	267(242)	267(269)
A-U	192(204)	380(402)
A-C	215(218)	262(277)
A-A	222(195)	388(308)
A-G	204(217)	255(300)
G-U	230(227)	326(391)
G-C	232(242)	346(269)
G-A	217(217)	327(300)
G-G	248(240)	252(291)
Σ	3568(3567)	5369(5372)

In Klammern sind die Erwartungswerte für die einzelnen 2er Sequenzen angegeben, die aus der durch Zählung ermittelten Basenzusammensetzung errechnet wurden.

Bei Gleichverteilung der Basen ergäbe sich für MS2 für jede 2er Sequenz ein Erwartungswert von 223 und für φX174 von 337.
Es muss nun getestet werden, inwiefern die gezählten Häufigkeiten der einzelnen Sequenzen signifikant von den Erwartungswerten abweichen. Die Wahrscheinlichkeit der Bildung einer bestimmten Sequenz aus den Basen N_1 und N_2 ist gleich: $pN_1 \cdot pN_2$ Hierfür kann man kurz pN_1N_2 schreiben. Insgesamt ergeben sich bei einer RNS der Länge M genau M-1 2er Sequenzen; der Erwartungswert für die Anzahl der auftretenden Sequenzen N_1-N_2 ist also gleich:

$(M-1) \cdot pN_1N_2$

Die gezählten Häufigkeiten einer Sequenz N_1-N_2 in allen möglichen RNS-Ketten der Länge M verhielten sich entsprechend einer Binomialverteilung. Ist M gross, so kann man diese Binomialverteilung ohne nennenswerten Fehler mit der Normalverteilung gleichsetzen und erhält damit die Möglichkeit vermittels des u-Tests ferzustellen, ob eine durch Zählung ermittelte Häufigkeit von der erwarteten Häufigkeit signifikant abweicht. Als signifikante Abweichung wird in der Folge ein Wert angesehen, der ausserhalb des 95% aller Werte umfassenden, symmetrischen Mittelteiles der Normalverteilung fällt (die oberen und unteren 2.5% der Verteilung fallen heraus). In der nächsten Tabelle sind für alle 16 möglichen 2er Sequenzen die zugehörigen, durch u-Test ermittelten, Integrale der Normalverteilung (Pr(u)) aufgeführt. Jeder Wert, der grösser als 0.975 ist, kann als eine signifikante Steigerung angesehen werden . Ist der Wert kleiner als 0.025, ist er signifikant gesunken.

	MS2	X174
U-U	0.4162	0.9808
U-C	0.9899	0.0146
U-A	0.1935	0.0000016
U-G	0.0851	0.9999959

	MS2	ϕX174
C-U	0.8127	0.9594
C-C	0.0730	0.0682
C-A	0.1174	0.1334
C-G	0.8971	0.5530
A-U	0.1935	0.7331
A-C	0.4266	0.8226
A-A	0.9766	0.9999980
A-G	0.1468	0.0037
G-U	0.5815	0.0003204
G-C	0.2528	0.0749
G-A	0.5000	0.9475
G-G	0.7036	0.0095

Bei MS2 wären somit U-C und A-A signifikant gestiegen und kein Wert signifikant gesunken. Bei X174 liegt bei U-U,U-G und A-A eine Steigerung und bei U-C,U-A,A-G,G-U und G-G eine Verminderung vor. Nur die Steigerung bei A-A ist beiden Phagen gemeinsam, und die Steigerung von U-C bei MS2 steht im Gegensatz zur Verminderung derselben Sequenz bei ϕX174.

Welche 2er Sequenzen müssten nun gesteigert und welche müssten vermindert sein, um eine vermehrte Sekundärstrukturbildung hervorzurufen? Dazu muss man sich zunächst einmal überlegen, welche der Sequenzen zueinander in Bindung treten könnten um ein Teilstück einer Doppelhelix zu bilden. Z.B. könnten U-U und A-A zusammentreten zu:

-U-U-
-A-A-

Eine Liste aller möglichen Paarbildungen zwischen 2er Sequenzen ist in der folgenden Tabelle gegeben:

U-A / U-A
U-G / U-G
C-G / C-G
A-U / A-U
G-U / G-U
G-C / G-C

U-U / A-A	A-A / U-U
U-U / A-G	A-G / U-U
U-U / G-A	G-A / U-U
U-U / G-G	G-G / U-U
U-C / G-A	G-A / U-C
U-C / G-G	G-G / U-C
U-G / C-A	C-A / U-G
U-G / C-G	C-G / U-G
C-U / A-G	A-G / C-U
C-U / G-G	G-G / C-U
C-C / G-G	G-G / C-C
A-U / G-U	G-U / A-U
A-C / G-U	G-U / A-C
G-U / G-C	G-C / G-U
U-A / U-G	U-G / U-G

Bei 16 2er Sequenzen ergeben sich 256 verschiedene Kombinationen, von denen aber nur 36 eine Basenpaarbindung aufweisen. Die Sequenzen U-U, U-G, G-U und G-G können in vier verschiedenen Kombinationen auftreten und die Sequenzen C-C, C-A, A-C und A-A können nur eine Art von Kombination bilden. Es wäre somit im Sinne einer erhöhten Fähigkeit zur Bildung von Sekundärstruktur, wenn die Häufigkeit der ersten gesteigert und die der letzten gesenkt wäre.
Von einer Veränderung der 2er-Sequenz-Häufigkeit im Sinne einer erhöhten Fähigkeit zur Sekundärstrukturbildung lässt sich also nur bei den Sequenzen U-U und U-G von φX174 sprechen. Jedoch stehen dem die signifikanten Veränderungen der Häufigkeiten von A-A bei MS2 und φX174 sowie von G-U und G-G bei φX174 in umgekehrter Richtung entgegen. Eine Anpassung der Primärsequenz an eine Notwendigkeit zur Erhöhung der Häufigkeit von komplementären 2er Sequenzen lässt sich auf diese Weise allerdings nicht erkennen.
Aus den Bindungsmöglichkeiten, die in der letzten Tabelle aufgeführt wurden, errechnet sich die Gesamthäufigkeit von komplementären 2er Sequenzen, die mit pBindung bezeichnet wird.

$$pBindung = pUA^2 + pUG^2 + pCG^2 + pAU^2 + pGU^2 + pGC^2 +$$
$$2pUGpUA + 2pCApUG + 2pCGpUG + 2pAGpCU +$$
$$2pAApUU + 2pAGpUU + 2pGApUU + 2pGGpUU +$$
$$2pGApUC + 2pGGpUC + 2pGUpAU + 2pGUpAC +$$
$$2pGGpCU + 2pGGpCC + 2pGCpGU$$

Wenn man definiert, dass $pN_1N_2N_3N_4 = pN_1N_2pN_3N_4$ ist, dann vereinfacht sich diese Formel zu:

$$pBindung = 4pUUAA + 4pUUGG + 4pCCGG +$$
$$8pUUGA + 8pACUG + 8pCGUG$$

Dies gilt allerdings nur, wenn $pN_1N_2 = pN_2N_1$ ist, d.h. wenn man in der Formel die pNN-Werte einsetzt, die sich nach den pN-Werten gemäss der Basenzusammensetzung errechnet haben. Unter dieser Vorraussetzung erhält pBindung bei MS2 den Wert 0.14249 und bei φX174 den Wert 0.15636. Bei Zugrundelegung der Häufigkeiten, die sich nach der realen Zählung ergaben, weist MS2 eine pBindung von 0.14277 auf und φX174 von 0.16361 . Die pBindung-Werte nach Schätzung entsprechend der Basenzusammensetzung und nach Zählung unterscheiden sich bei MS2 nicht signifikant (u-Test: u = 1.079 , Pr(u) = 0.8560), jedoch ist dies bei φX174 der Fall (u-Test: u = 42.362, Pr(u) = 1.000) . Nur bei φX174 konnte also eine signifikante Anpassung der Verteilung der 2er Sequenzen an eine erhöhte Bildung von Sekundärstrukturen festgestellt werden.

Es fragt sich als Nächstes, ob die Veränderungen der Häufigkeiten von 2er Sequenzen alle möglichen Kombinationen mit Basenpaarbindungen fördern, oder ob bestimmte Kombinationen bevorzugt gesteigert werden. Es wäre zum Beispiel denkbar, dass die Kombinationen, die nur aus (G·C)-Basenpaaren bestehen mehr gefördert würden als andere, da aus thermodynamischen Experimenten bekannt ist, dass (G·C)-Basenpaare stabiler sind als (A·U)-Basenpaare und insbesondere (G·U)-Basenpaare.

Es sollen nun diejenigen Kombinationen, die nur aus (G·C)-Basenpaaren bestehen, mit denen verglichen werden, die nur aus (G·U)-Basenpaaren bestehen. Die Ersten sind als die thermodynamisch stabilsten anzusehen, den Letzten kann man keine nennenswerte Stabilität zuordnen. Die stabilen Kombinationen haben nach Tabelle folgende Wahrscheinlichkeit(pStabil):

$$pStabil = 2pGGpCC + pCG^2 + pGC^2$$

Diese Formel vereinfacht sich wiederum für den Fall, dass man die aus der Basenzusammensetzung errechneten pNN-Werte benutzt und somit $pN_1N_2 = pN_2N_1$ gilt zu:

$$pStabil = 4pGGCC$$

Die instabilen Kombinationen erhalten die Wahrscheinlichkeit pInstabil:

$$pInstabil = 2pGGpUU + pGU^2 + pUG^2$$

Dies vereinfacht sich unter denselben Voraussetzungen wie oben zu:

$$pInstabil = 4pGGUU$$

Nach diesen Formeln ergeben sich die Werte der folgenden Tabelle:

	MS2	φX174
pStabil(nach Basenzus.)	$1.8389 \cdot 10^{-2}$	$1.0095 \cdot 10^{-2}$
pStabil(nach Zählung)	$1.8506 \cdot 10^{-2}$	$1.0595 \cdot 10^{-2}$
pInstabil(nach Basenzus.)	$1.8389 \cdot 10^{-2}$	$2.1119 \cdot 10^{-2}$
pInstabil(nach Zählung)	$1.5733 \cdot 10^{-2}$	$2.1478 \cdot 10^{-2}$

Zum Vergleich dieser Zahlen wurde wieder ein u-Test durchgeführt:

Vergleich	MS2	φX174
pStabil	u = 0.4188 Pr(u) = 0.6623	u = 2.6981 Pr(u) = 0.9965
pInstabil	u = -9.5686 Pr(u) = 0.000	u = 1.9488 Pr(u) = 0.9743

Aus diesen Zahlen wird ersichtlich, dass bei MS2 trotz einer nicht signifikanten Veränderung der Häufigkeit stabiler Kombinationen doch das Auftreten instabiler mit hoher Signifikanz vermindert ist und bei φX174 eine signifikante Erhöhung der stabilen Elemente auftritt, die mit einer nicht signifikanten Veränderung der instabilen verbunden ist. Sämtliche Veränderungen sind also auch qualitativer Art, es werden stabile Sekundärstrukturen eher gefördert als instabile, auch wenn, wie im Falle von MS2, keine signifikante Erhöhung der Gesamtmenge an Kombinationen von 2er Sequenzen auftritt.

Zusammenfassend lässt sich sagen, dass durch die Veränderung der Sequenz des RNS-Moleküles bei MS2 und φX174 sowohl eine leichte Steigerung der Häufigkeit der Bildung von Basenpaaren bei Kontakt zweier beliebiger 2er Sequenzen auftritt als auch eine qualitative Steigerung von stabilen Basenpaarbildungen gegenüber instabilen stattfindet. Eine weitergehende Analyse könnte dasselbe Verfahren, auf die Häufigkeiten von 3er,4er,5er etc Sequenzen anwenden. Dies führt allerdings wegen der Kürze der bekannten Sequenzen nicht zu statistisch signifikanten Werten. Bereits bei den 3er Sequenzen sind nicht alle möglichen in MS2 vertreten. Eine einfachere Art,die Anpassung längerkettiger Sequenzen an die Notwendigkeit zur Bildung von Sekundärstrukturen zu testen,ist durch Auszählung längerkettiger helicaler Bereiche in der RNS gegeben, die in Abschnitt 4.2 durchgeführt wird.

4.1.6 Interpretation als MARKOFF-Kette

Nach der Untersuchung der Basenzusammensetzung wurde gefragt, inwieweit die Veränderung der Basenzusammensetzung zu einer Verminderung der Fähigkeit der RNS, als genetisches Material zu fungieren, führt. Wird die Informationskapazität der RNS aber nicht auch durch eine veränderte Häufigkeit von 2er Sequenzen verringert? Zur Beantwortung dieser Frage muss nun ein neues Informationsmass angewendet werden, dass auf einer Interpretation der RNS als MARKOFF-Kette basiert.

Eine MARKOFF-Kette ist eine unendliche Zeichenreihe, von der allein bekannt ist, mit welcher Wahrscheinlichkeit irdendeines der Zeichen auftritt, wenn zuvor zuvor irgendein anderes Zeichen aufgetreten ist. Dies nennt man Transitionswahrscheinlichkeit. Aus der Tabelle der Häufigkeiten der einzelnen 2er Sequenzen lassen sich nun Schätzungen für die Wahrscheinlichkeiten errechnen, mit denen die vier Zeichen der RNS aufeinanderfolgen. Aus diesen Transitionswahrscheinlichkeiten lässt sich der mittlere Informationsgehalt eines Zeichens errechnen.

Die Wahrscheinlichkeit eines bestimmten Überganges oder einer bestimmten Aufeinanderfolge (Transition) von zwei Basen wird mit $p(N_1 \rightarrow N_2)$ bezeichnet. Somit sind:

$$p(U \rightarrow U) = \frac{pUU}{pUU + pUC + pUA + pUG}$$

$$p(U \rightarrow C) = \frac{pUC}{pUU + pUC + pUA + pUG}$$

$$p(U \rightarrow A) = \frac{pUA}{pUU + pUC + pUA + pUG}$$

$$p(U \rightarrow G) = \frac{pUG}{pUU + pUC + pUA + pUG}$$

Und mit $pU = pUU + pUC + pUA + pUG$ ergibt sich:

$$p(U \rightarrow U) = \frac{pUC}{pU}, \quad p(U \rightarrow C) = \frac{pUC}{pU}, \quad p(U \rightarrow A) = \frac{pUA}{pU},$$

$$p(U \rightarrow G) = \frac{pUG}{pU}.$$

und entsprechend für die Übergänge p(C→N), p(A→N) und p(G→N). Die Entropie oder der Informationsgehalt eines Überganges(I(N→)) wird entsprechend der oben aufgeführten Formel für die Information einer beliebigen Zeichenkette berechnet (siehe ASHBY,S.257-258 (1a)):

$$I(U\rightarrow) = - p(U\rightarrow U)ld(p(U\rightarrow U)) - p(U\rightarrow C)ld(p(U\rightarrow C)) - p(U\rightarrow A)ld(p(U\rightarrow A)) - p(U\rightarrow G)ld(p(U\rightarrow G)).$$

$$I(C\rightarrow) = - p(C\rightarrow U)ld(p(C\rightarrow U)) - p(C\rightarrow C)ld(p(C\rightarrow C)) - p(C\rightarrow A)ld(p(C\rightarrow A)) - p(C\rightarrow G)ld(p(C\rightarrow G)).$$

$$I(A\rightarrow) = - p(A\rightarrow U)ld(p(A\rightarrow U)) - p(A\rightarrow C)ld(p(A\rightarrow C)) - p(A\rightarrow A)ld(p(A\rightarrow A)) - p(A\rightarrow G)ld(p(A\rightarrow G)).$$

$$I(G\rightarrow) = - p(G\rightarrow U)ld(p(G\rightarrow U)) - p(G\rightarrow C)ld(p(G\rightarrow C)) - p(G\rightarrow A)ld(p(G\rightarrow A)) - p(G\rightarrow G)ld(p(G\rightarrow G)).$$

Der mittlere Informationsgehalt pro Zeichen (I(→)) wird dann berechnet unter Berücksichtigung der relativen Häufigkeit der einzelnen Zeichen, d.h. in unserem Fall der Basenzusammensetzung:

$$I(\rightarrow) = I(U\rightarrow)pU + I(C\rightarrow)pC + I(A\rightarrow)pA + I(G\rightarrow)pG$$

Um festzustellen, ob mit der Veränderung der 2er Sequenzhäufigkeiten eine einschneidende Veränderung der Informationskapazität der RNS verbunden ist, muss I(→) zum einen unter Benutzung der, mittels der Basenzusammensetzung ermittelten Häufigkeiten pNN errechnet werden und zum anderen unter zuhilfenahme der durch Zählung ermittelten pNN-Werte. In der nächsten Tabelle sind die Werte der p(N→N) für MS2 und ϕX174 aufgeführt (in Klammern sind diejenigen Werte, die sich nach Schätzung aus der Basenzusammensetzung ergeben):

	MS2	ϕX174
p(U→U)	0.2411(0.2449)	0.3387(0.3121)
p(U→C)	0.3006(0.2630)	0.1914(0.2154)
p(U→A)	0.2217(0.2334)	0.1860(0.2399)

	MS2	φX174
p(U→G)	0.2366(0.2597)	0.2838(0.2333)
p(C→U)	0.2594(0.2454)	0.3486(0.3120)
p(C→C)	0.2390(0.2615)	0.1964(0.2161)
p(C→A)	0.2151(0.2337)	0.2240(0.2394)
p(C→G)	0.2862(0.2594)	0.2310(0.2325)
p(A→U)	0.2305(0.2446)	0.2957(0.3124)
p(A→C)	0.2581(0.2614)	0.2039(0.2152)
p(A→A)	0.2665(0.2338)	0.3019(0.2393)
p(A→G)	0.2449(0.2602)	0.2984(0.2331)
p(G→U)	0.2481(0.2451)	0.2606(0.3125)
p(G→C)	0.2503(0.2613)	0.2766(0.2150)
p(G→A)	0.2341(0.2343)	0.2614(0.2398)
P(G→G)	0.2675(0.2592)	0.2014(0.2326)

Hieraus errechnen sich sodann die I(N→)-Werte:

	MS2	φX174	
I(U→)	1.9899(1.9652)	1.9527(1.9853)	bit
I(C→)	1.9922(1.9985)	1.9630(1.9850)	bit
I(A→)	1.9978(1.9985)	1.9722(1.9848)	bit
I(G→)	1.9984(1.9986)	1.9900(1.9846)	bit

und auch I(→) :

	MS2	φX174
I(→)	1.9946(1.9904)	1.9683(1.9846)

Es zeigte sich, dass bei MS2 die Abweichungen der p(N→N)-Werte von dem Wert 0.2500, der bei Gleichverteilung zu erwarten wäre, geringer sind als bei φX174. Dies spiegelt sich auch in den einzelnen Informationskapazitäten wider, die den Transitionswahrscheinlichkeiten der einzelnen Basen zugeordnet sind, sie sind bei φX174 allgemein geringer als bei MS2. Der Verlust an Informationskapazität gegenüber dem Maximalwert von 2.0 ist bei MS2 numerisch kaum erfassbar (kleiner als 0.01) und der Verlust gegenüber

Erwartungswert von 1.9904 ist durch den obligaten Rechenfehler verwischt worden. Bei ΦX174 zeigt sich ein Informationskapazitätsverlust der kleiner als 0.05 bit ist, und der Verlust gegenüber dem Erwartungswert ist geringer als 0.01 bit.

Es fragt sich nun, ob diese genannten Informationsverluste eine ernstzunehmende Beeinträchtigung der Codierfähigkeit des Moleküls darstellen. Von H.P. YOCKEY (40) wurde errechnet, dass die Informationskapazität, die benötigt wird, um ein Protein zu kodieren unter Berücksichtigung der Codedegeneration und der Tatsache, dass Aminosäuren mit gleichem Verhalten gegen Wasser (hydrophil,hydrophob) zu einem gewissen Ausmass gegeneinander austauschbar sind, pro Base nur etwa 1.347 bit beträgt. Allein deshalb kann der oben geschilderte Verlust nur gering sein, durch das Auftreten von Genüberlappungen wie in ΦX174 hat sich fernerhin gezeigt, dass Proteine auch mit noch geringerer Informationskapazität der Nucleinsäure (in diesem Fall 1.0 bit pro Base) sinnvoll codiert werden können.

4.2 Helixzählung

Es wurde versucht, zu einem Verfahren zu gelangen, mit dessen Hilfe es gelingt, sämtliche möglichen helicalen Bereiche einer einsträngigen RNS aufzufinden, statistisch auszuwerten und für eine spätere Verwendung bei der Untersuchung der Gesamtsekundärstruktur der RNS zu speichern.

4.2.1 Bindungsmatrix (base pair matrix)

Um sich über bestimmte Bereiche der RNS konkret verständigen zu können, hat man sämtliche Nucleotide, angefangen vom 5'-Ende, fortlaufend numeriert. Wenn nun eine WATSON-CRICK- oder Wobble-Bindung zwischen zwei Basen desselben Moleküls vorkommt, so heisst dies, dass zwei unterschiedlich numerierte Basen, die einen gewissen Abstand voneinander haben, kombiniert werden. So ist es zum Beispiel möglich, dass sich die Base Nr. 9 (ein Cytosin) mit der Base Nr. 20 (ein Guanin) von MS2 kombiniert. Mathematisch gesehen stellen diese Kombinationen eine Abbildung eines Teiles der natürlichen Zahlen (von 1 bis 3569, der Länge von MS2) in sich selbst dar. Man kann diese Abbildung durch eine Matrix graphisch darstellen, deren Zeilen- und Spaltenzahl gleich 3569 ist und bei der jedes Matrixelement eine Kombination zweier numerierter Nucleotide repräsentiert. In Abb. 15 ist eine solche Matrix für die ersten 20 Nucleotide von MS2 dargestellt und es wurde in jedem Matrixelement eingetragen, ob entweder eine (A·U)-Bindung (Kreis), (G·C)-Bindung (Quadrat) oder (G·U)-Bindung (auf der Spitze stehendes Quadrat) vorliegt. Es zeigt sich, dass diese Matrix aller möglichen Bindungen (Bindungsmatrix oder auch Basenpaarmatrix (engl.:base pair matrix) gegenüber der Diagonale von Links oben nach Rechts unten spiegelsymmetrisch ist. Dies wird verständlich, wenn man sich überlegt, dass die Verknüpfung der Base n mit der Base m iden-

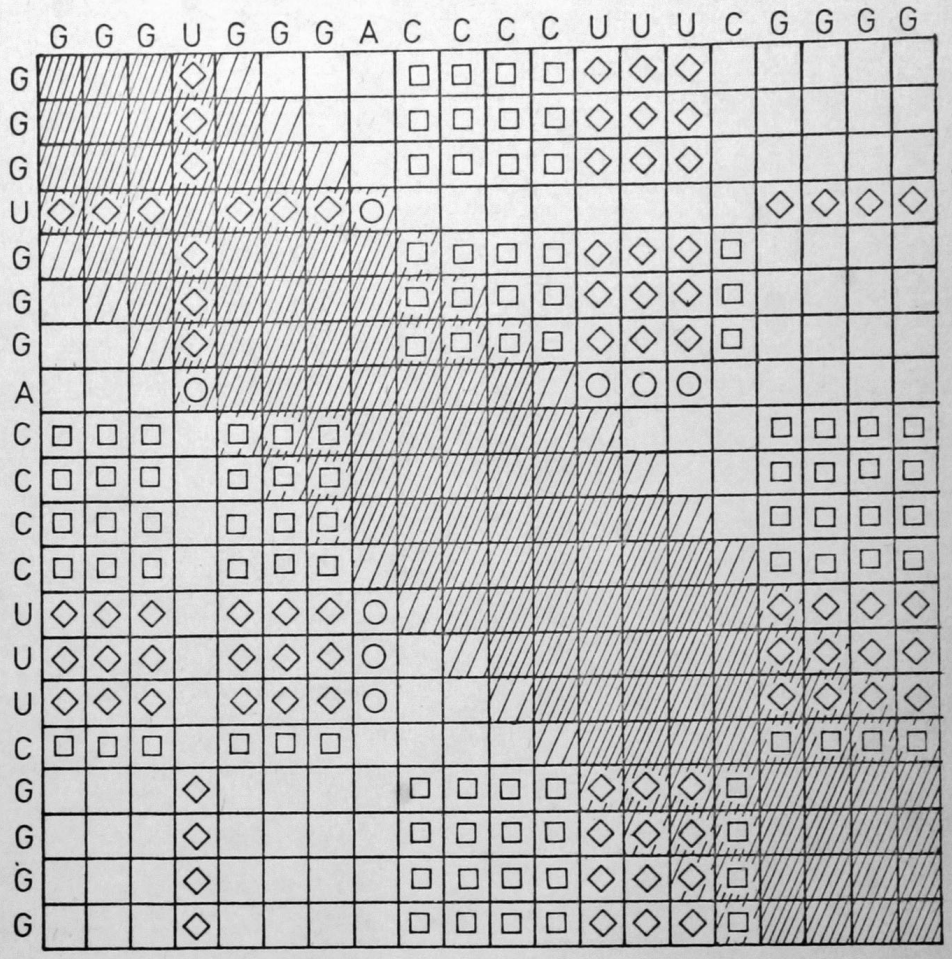


Abb. 15: Basenpaarmatrix der ersten 20 Nucleotide von MS2. Kreis = (A·U), Quadrat = (G·C), Auf der Spitze stehendes Quadrat = (G·U)

tisch sein muss mit der Verknüpfung der Base m mit der Base n. Die Matrixelemente n,m und m,n sind aber symmetrisch gegenüber der genannten Diagonale. Aus diesem Grunde kann für die folgenden Überlegungen die eine Hälfte der Matrix ausser Betracht gelassen werden, ich werde in Zukunft die obere Hälfte weglassen. Weiterhin können sehr nahe beieinanderliegende Basen nicht miteinander koppeln. Nach der Literatur (37) müssen zwischen den beiden Basen einer (G·C)-Bindung mindestens 3 Basen Zwischenraum sein und zwischen den beiden Basen einer (A·U)-Bindung mindestens 4 ungebundene Nucleotide. Für (G·U)-Bindungen sind derartige Abstände

nicht bekannt, darum wurden (G·U)-Bindungen im Rahmen dieser Untersuchungen (A·U)-Bindungen in Bezug auf ihre Mindestabstände gleichgesetzt. Diese "zu engen" Bindungen liegen nun entlang der oben genannten Diagonale, was man sich leicht klarmachen kann, wenn man bedenkt, dass der Abstand zweier Basen gleich der Differenz ihrer fortlaufenden Nummern minus eins ist. Diese Differen ist desto geringer, je näher eine, der Diagonalen paralle Linie an diese heranrückt (in der Diagonalen selbst ist die Differenz gleich -1, da n=m). In der obigen Abbildung brauchen also nur die Basenpaare im nicht schraffierten Bereich berücksichtigt zu werden, und im weniger stark Schraffierten Bereich brauchen nur die (G·C) Basenpaare beachtet zu werden.
Eine Helix ist eine fortlaufende Kette von Basenpaaren und in der Matrix stellt sich diese Kette als eine fortlaufende Reihe von Matrixelementen dar, die parallel zur zweiten Diagonale (von Links unten nach Rechts oben) angeordnet sind.
Abb. 16 zeigt eine mögliche Sekundärstruktur des 1-20-Abschnittes von MS2:

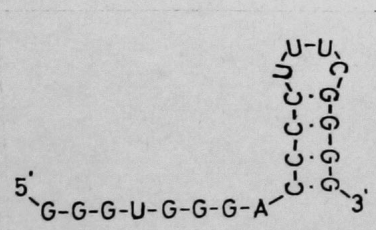


Abb. 16: Mögliche Sekundärstruktur der ersten 20 Nucleotide von MS2

Diese Helix, bei der die Elemente 9 bis 12 mit den Elementen 17 bis 20 verknüpft sind, findet sich in der Bindungsmatrix wieder. In Abb. 17 sind alle Basenpaare, die in eine mögliche Helix integriert werden können durch eine Linie verbunden und die Helix von Abb. 16 ist besonders hervorgehoben. Zum einen sind zu nah an der Diagonale liegende (G·U)-Paare nicht in die Helices integriert und zum anderen blei-

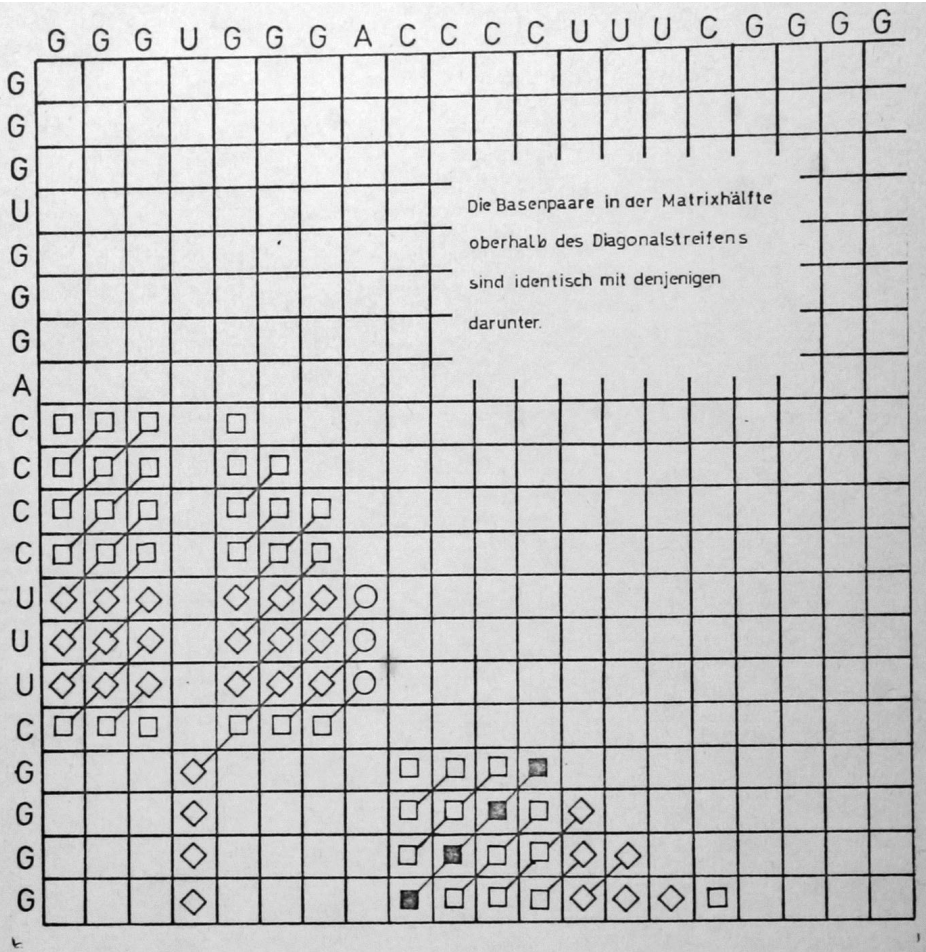


Abb. 17: Die zu einer Helix gehörigen Basenpaare sind mit einer Linie verbunden. Die Helix aus Abb.16 besitzt schwarz ausgefüllte Rechtecke.

ben einige vereinzelt liegende Basenpaare wie 4,18 und 3,16 unberücksichtigt. In der Folge werden Basenpaare durch Angabe der sie bildenden Basen und deren Nummern bezeichnet. Die Nummern werden als untere Indizes eingesetzt:

z.B. $(G_3 \cdot C_{16})$

oder, wenn nicht bekannt ist welche Basen auftreten:

$(N_3 \cdot N_{16})$

N steht für Nucleotid. Auch können die Zahlen durch variable Indizes ersetzt werden:

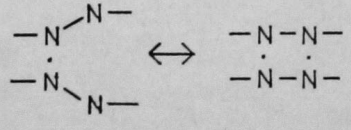
$(N_i \cdot N_j)$

Abb. 17 entnimmt man, dass es in der Sequenz 1-20 von MS2 genau 23 verschiedene mögliche helicale Bereiche oder Helices gibt. Da die Anzahl der möglichen Helices in etwa proportional zur Fläche der halben Matrix ist und diese Fläche proportional mit dem Quadrat der Länge des betrachteten RNS-Moleküles wächst, wächst auch die Anzahl der möglichen Helices quadratisch mit der Länge der RNS.

Eine strukturell gleiche Matrix wurde ebenfalls von TINOCO et al. (1971) (38) zur Betrachtung von möglichen Sekundärstrukturen entwickelt und 1974 von GRALLA und DeLISI (15) verwendet. Auch PIPAS und McMAHON sprechen 1975 von einer "bonding matrix" (32), die vermutlich der hier erläuterten ähnlich ist.

4.2.2 Berechnung der freien Energie einer RNS-Helix

Es ist von verschiedener Seite versucht worden RNS-Doppelhelices in Abhängigkeit von ihrer Sequenz eine freie Energie (ΔG) zuzuordnen. Am weitesten gedieh dieser Versuch in einem Artikel aus dem Jahre 1973 von TINOCO et al.(37), in dem sowohl bestimmten Basenpaarfolgen wie auch den zwischen den gebundenen Abschnitten liegenden freien Schleifen und Enden des Moleküls eine entsprechende freie Energie zugeordnet wird. Die freie Energie von Helices hat negative Werte, je kleiner also diese freie Energie ist, desto grösser wird die Stabilität der Helix. Die Stabilität verbessert sich also, wenn die freie Energie abnimmt. Umgekehrt liegt der Fall bei den ungebundenen Abschnitten, diese destabilisieren das Molekül und ihnen werden positive freie Energien zugeordnet. Es ist nicht möglich, einfach einem bestimmten Basenpaar eine freie Energie zuzuordnen, denn ein grosser Teil der Stabilität einer Doppelhelix resultiert aus den Interaktionen zwischen übereinanderliegenden Basenpaaren. Es kann nur versucht werden, 2 nebeneinanderliegenden Basenpaaren eine freie Energie zuzusprechen, die sich zusammensetzt aus den Bindungsenergien zwischen den Basen des Paares, die vorwiegend polarer Natur sind, und den vorwiegend unpolaren Bindungsenergien zwischen den Basenpaaren. Diese zusammengesetzte freie Energie wird weiterhin als Stapelkraft bezeichnet, in Anlehnung an die englische Bezeichnung "stacking force". Die Stapelkraft ist als die freie Energie (ΔG) der folgenden Reaktion definiert:



In Abb. 18 sind die freien Energien aller möglichen 2er Helices in kcal angegeben, wie sie der Arbeit von TINOCO et al. (37) entnommen wurden. Es werden auch solche 2er Helices aufgeführt, die (G·U)-Basenpaare enthalten. Waagrecht ist das erste Basenpaar tabelliert und senkrecht das zweite.

		1·2 U·A	1·2 A·U	1·2 G·C	1·2 C·G	1·2 G·U	1·2 U·G
3·4	U·A	-1.2	-1.8	-2.2	-2.2	0.0	0.0
3·4	A·U	-1.8	-1.2	-2.2	-2.2	0.0	0.0
3·4	G·C	-2.2	-2.2	-5.0	-3.2	0.0	0.0
3·4	C·G	-2.2	-2.2	-5.0	-5.0	0.0	0.0
3·4	G·U	0.0	0.0	0.0	0.0	0.0	-0.3
3·4	U·G	0.0	0.0	0.0	0.0	-0.3	0.0

Abb. 18: Stapelkräfte nach TINOCO et al. (37) in kcal

Da die Stapelkräfte in kcal angegeben sind, wurden in weiteren Bearbeitungen, insbesondere in den Computerprogrammen kcal verwendet anstelle der heute üblichen kJ. Dies hat den Vorteil, dass die Programme mit Werten rechnen, die auf eine Stelle hinter dem Komma rundbar sind. Programmergebnisse können durch Multiplikation mit dem Faktor 4.1855 in kJ umgerechnet werden. Nach der Tabelle in Abb. 18 sind alle weiteren Werte für die Stabilität von helicalen Bereichen in dieser Arbeit errechnet worden. Es gibt noch einen neueren Versuch diese thermodynamischen Werte zusammenzustellen (BORER et al. (1974) (4)), jedoch ist diese Liste nicht vollständig, es fehlen Werte für die in Abschnitt 4.3 näher definierten ungebundenen Teilsequenzen, die

ebenfalls in der Arbeit von TINOCO et al. enthalten sind. Die Werte in der Arbeit von BORER et al. zeigen weiterhin keine prinzipiellen Unterschiede, sie liegen allgemein etwas niedriger, stehen aber zueinander in etwa in denselben Verhältnissen. Auch sie seien der Vollständigkeit halber in Abb. 19 wiedergegeben.

	U·A	A·U	G·C	C·G	G·U	U·G
U·A		-1.6				
A·U		-1.2	-2.1			
G·C		-2.1		-3.0		
C·G			-4.3	-4.8		
G·U						
U·G						

Abb. 19: Stapelkräfte nach BORER et al.(4) in kcal

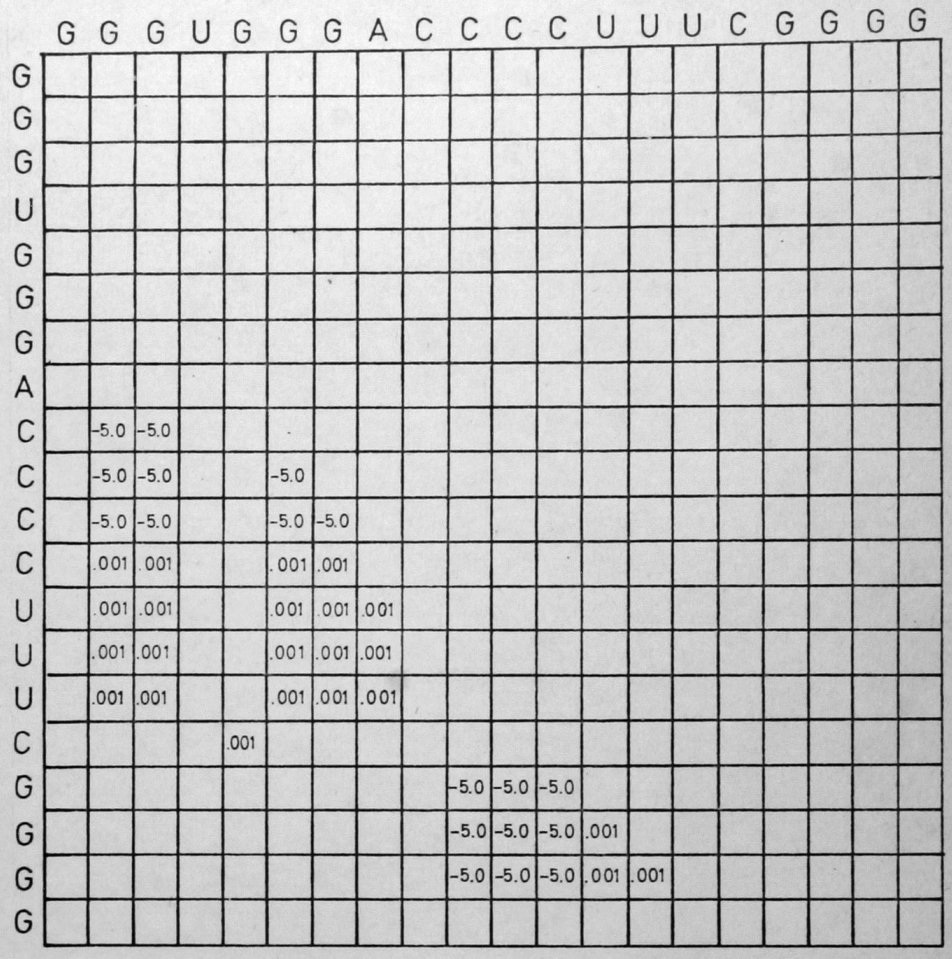
So hat die als Beispiel in Abschnitt 4.2.1 genannte mögliche Helix im Nucleotidbereich 1-20 von MS2 eine freie Enerige von -15.0 kcal oder -62.7825 kJ, weil in ihr dreimal die 2er Helix $\begin{matrix}-C-C-\\ \vdots\ \vdots\\ -G-G-\end{matrix}$ vorkommt, deren Stabilität oder freie Energie gleich -5.0 kcal (-20.9275 kJ) ist.

4.2.3 Beschreibung der Computerprogramme zur Gewinnung aller möglichen Helices einer RNS und Berechnung ihrer Stabilitäten

Um die Bindungsmatrix einer beliebigen RNS automatisch nach sämtlichen möglichen Helices durchsuchen zu lassen, damit man diese für statistische Zwecke und als Ausgangsbasis für Untersuchungen der möglichen Konformationen des Gesamtmoleküls verwenden kann, wurden zwei Computerprogramme entwickelt, die das Problem auf unterschiedliche Weise lösen. Das 1./ Programm, geschrieben in der Programmiersprache FORTRAN IV, ist Teil des willkürlich FALTUNG genannten Programmes und basiert auf der konkreten Speicherung der Bindungsmatrix im Computer. Das zweite Programm, das nur mit einer virtuellen Speicherung der Bindungsmatrix arbeitet, wurde in der Sprache SIMULA erstellt und HELIX. LISTE genannt.

4.2.3.1 Errechnung der möglichen Helices im Programm FALTUNG

Die Bindungsmatrix ist programmintern durch einen 2-dimensionalen Vektor $B_{i,j}$ dargestellt. Da man den einzelnen Basenpaaren keine freie Energie zuordnen kann, wird das ΔG der 2er Helix aus den Basenpaaren $(N_i \cdot N_j)$ und $(N_{i+1} \cdot N_{j-1})$ an der Stelle i+1,j-1 in die Matrix B eingetragen. Als Beispiel seien für den Bereich 1-20 von MS2 die entsprechenden freien Energien wiedergegeben(Abb.20). Wo keine Bindung möglich ist oder nur vereinzelte Basenpaare auftreten, wird in die Matrix der Grundwert 0.0 eingetragen. Um diesen Grundwert von Fällen zu unterscheiden, wo zwar eine Basenpaarbindung auftritt, aber die freie Energie gleich 0.0 ist, wird der numerisch nicht ins Gewicht fallende Wert 0.001 eingetragen. In welcher Reihenfolge das Programm die Stabilitätswerte in die Matrix einträgt, ist hierbei ohne Belang. Anders wird



	G	G	G	U	G	G	G	A	C	C	C	C	U	U	U	C	G	G	G	G
G																				
G																				
G																				
U																				
G																				
G																				
G																				
A																				
C		-5.0	-5.0																	
C		-5.0	-5.0			-5.0														
C		-5.0	-5.0			-5.0	-5.0													
C		.001	.001			.001	.001													
U		.001	.001			.001	.001	.001												
U		.001	.001			.001	.001	.001												
U		.001	.001			.001	.001	.001												
C					.001															
G										-5.0	-5.0	-5.0								
G										-5.0	-5.0	-5.0	.001							
G										-5.0	-5.0	-5.0	.001	.001						
G																				

Abb. 20: Matrix B des Programms FALTUNG mit den freien Energien (in kcal) aller möglichen 2er Helices für das Beispiel der Nucleotide 1-20 von MS2. Grundwerte 0.0 weggelassen.

dies bei der Aufsummierung dieser Werte, um die Gesamtstabilitäten der Helices aufzufinden. In der Matrix B wird in der Weise, wie dies in Abb. 21 geschehen ist, jeweils der Speicherinhalt des Matrixelementes i,j mit dem Speicherinhalt des Elementes i-1,j+1 addiert und in i,j eingetragen. Dies erfolgt spaltenweise von links nach rechts und innerhalb der Spalten von oben nach unten. Hat der Speicherinhalt an der Stelle i,j den Grundwert 0.0, dann erfolgt keine Addition. Es werden also bildlich gesehen die freien Energien einer Helix von links unten nach rechts oben aufaddiert und die Gesamt-freie

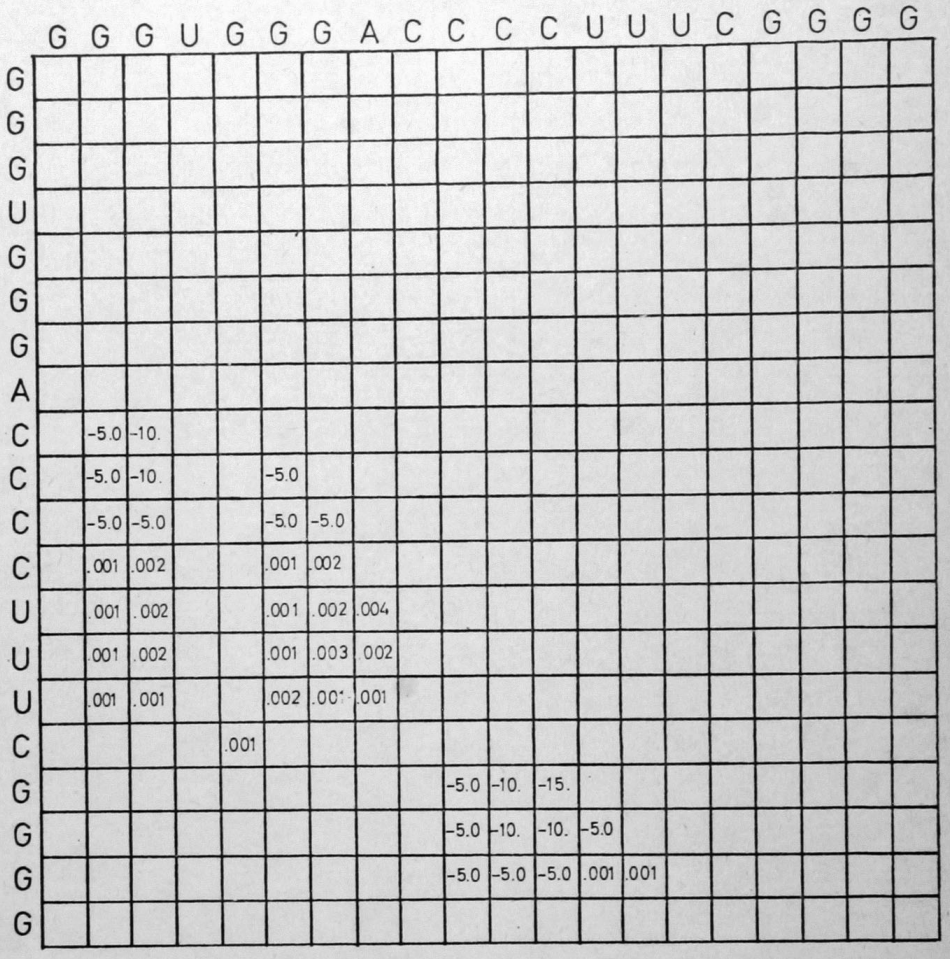


Abb. 21: Auf_addierung der Freien Energien (in kcal) der Helices in Abschnitt 1-20 von MS2 entsprechend dem Programm FALTUNG. Die Auf-addierung erfolgt von links unten nach rechts oben. Z.T. sind gerundete Werte eingetragen.

Energie der Helix sammelt sich am rechten oberen Ende der Helixrepräsentation.
Dieses Verfahren benötigt sehr viel Speicherplatz, weil die Grösse des Vektors B mit der Länge des untersuchten RNS-Abschnittes quadratisch wachst. Es hat sich herausgestellt, dass dieses Programm nur bei RNS-Molekülen oder Teilsequenzen bis zu einer Länge von etwa 300 Nucleotiden ökonomisch eingesetzt werden kann.

4.2.3.2 Errechnung der möglichen Helices im Programm HELIX.LISTE bzw. ZAEHLUNG

Wegen des hohen Speicherplatzbedarfes ist das Programm FALTUNG unrentabel für längerkettige RNS-Moleküle oder Teilsequenzen. Von der Matrix B, deren sämtliche Elemente in der Rechenmaschine Speicherplatz beanspruchen, wird wegen der Symmetrie der Bindungsmatrix gegenüber ihrer Diagonale nur die Hälfte benötigt. Auch ist sodann der grösste Teil dieser Hälfte mit dem Grundwert 0.0 belegt, der für die weitere Rechnung nicht benötigt wird. Es wurde darum versucht, in zwei weiteren Programmen in der Programmiersprache SIMULA nur die zum Erkennen einer Helix notwendigsten Daten in den Speicher aufzunehmen. Um eine Helix exakt beschreiben zu können, sind nur drei ganze Zahlen notwendig: zum einen handelt es sich um die i und j Positionen des Basenpaares, das an einem Ende der Helix liegt und sodann um dieGesamtzahl der in der Helix verknüpften Basenpaare, also der Länge der Helix. Es wurde vereinbart, immer nur dasjenige Ende der Helix zu speichern, dass, in der Matrix gesehen, links unten liegt. Mit Hilfe der Länge [illegible] der Helix lassen sich sodann alle anderen Werte, z.B. die Koordinaten des Zweiten Endes, bestimmen; hat das linke untere Ende die Koordinaten i,j und besitzt die Helix die Länge n, so hat das rechte obere Ende die Koordinaten i+n-1,j-n+1 . HELIX.LISTE und ZAEHLUNG durchsuchen nun die Bindungsmatrix nach allen möglichen Helices und speichern diese durch die genannten Zahlentripel. Während der Suche wird die Matrix B nicht real im Speicher der Maschine abgebildet, sondern es wird nur noch virtuell mit dieser Matrix gearbeitet. Um die Anfangs- und Endkoordinaten einer Helix aufzufinden, lässt man die Indices i und j so varriieren, dass ihre Veränderung einer Wanderung in der Matrix nach dem in Abb. 22 wiedergegebenen Schema entspricht.

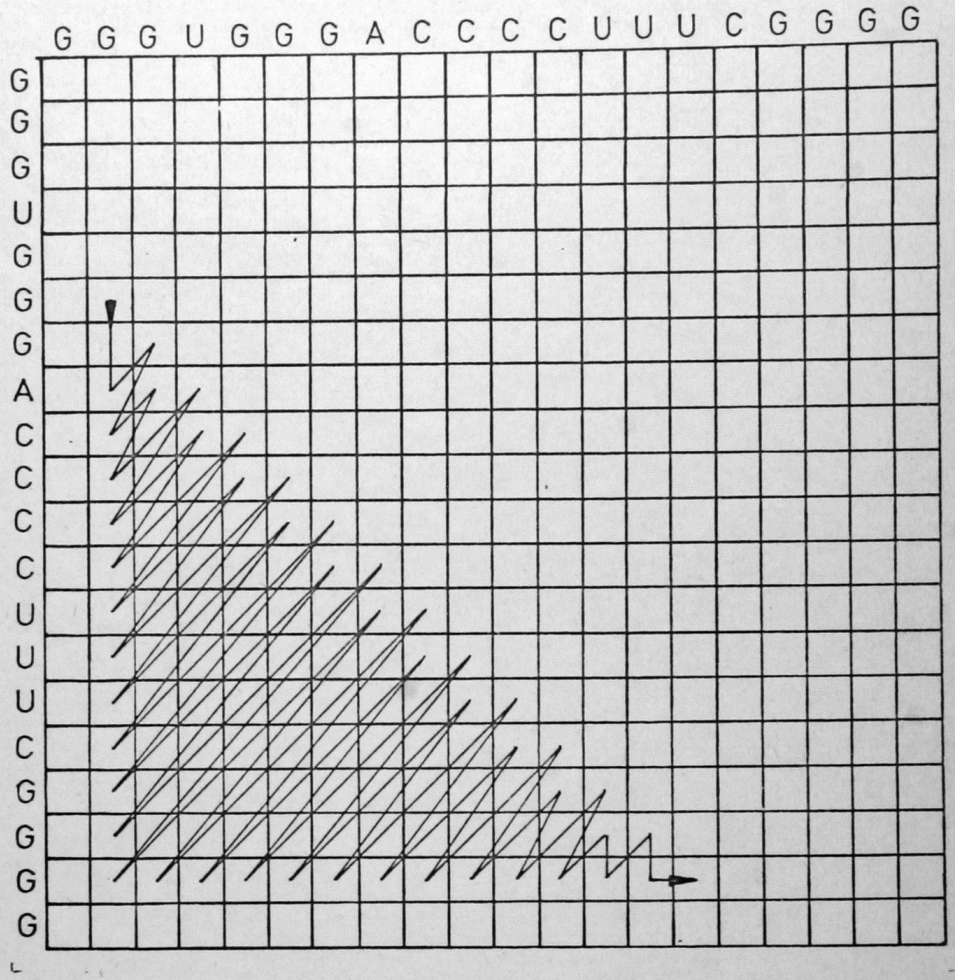


Abb. 22: Indexveränderung in den Programmen HELIX.LISTE und ZAEHLUNG am Beispiel des Abschnittes 1 bis 20 von MS2.

Die Elemente der Matrix werden also von links unten nach rechts oben abgetastet, und es kann anhand einer Abfrage an jedem Punkt der Matrix festgestellt werden, wann eine fortlaufende Kette von Basenpaarbindungen beginnt und endet. Beginnt eine solche Kette, werden die Koordinaten i und j als Anfangsstelle einer Helix notiert und dann wird wird ausgezählt, wie lange eine ununterbrochene Kette von Bindungen folgt und die Länge der Helix festgehalten. Gleichzeitig kann die Stabilität einer Helix errechnet werden und bei Bedarf ebenfalls abgespeichert werden, allerdings ist es auch möglich aus den Angaben über Anfang und Länge einer

Helix jederzeit nachträglich ihre Stabilität, ausgedrückt durch ihre freie Energie zu berechnen.

Die Programme ZAEHLUNG und HELIX.LISTE unterscheiden sich in der Art ihres outputs. Von ZAEHLUNG wird festgestellt, wie oft eine Helix mit einer bestimmten Sequenz von Basenpaaren in der Liste aller möglichen Helices vertreten ist. Es führt dann alle aufgetretenen unterschiedlichen Sequenzen auf und gibt an wie oft sie vorkommen. Die Sequenzen werden nach ihrer freien Energie und ihrer Länge geordnet. HELIX.LISTE ist variabel gehalten und kann je nach Eingabe unterschiedlicher Parameter entweder eine Liste mit einer Auszählung der Helices nach Länge und nach der Stabilität anfertigen oder druckt direkt alle Helices in Form von Zahlentripeln aus. Letzteres ist dann sinnvoll, wenn die Helices für Sekundärstruktur-Untersuchungen weiter verwendet werden sollen. Für das Beispiel der Teilsequenz 1-20 von MS2 wurde von ZAEHLUNG folgende Liste angefertigt:

Helixstatistik

Anzahl	ΔG	Sequenz
1	0.0	-G-G-A- : : : -C-U-U-
2	0.0	-G-G-G- : : : -U-U-U-
1	0.0	-G-G-G- : : : -C-U-U-
2	0.0	-G-G-G- : : : -U-U-C-
1	0.0	-U-G-G-G-A- : : : : : -G-C-U-U-U-
1	-2.2	-G-A- : : -C-U-
1	-5.0	-C-C- : : -G-G-
3	-5.0	-G-G- : : -C-C-

Anzahl	Δ G	Sequenz
1	-5.0	-C-C-U- : : : -G-G-G-
2	-5.0	-G-G-G- : : : -U-C-C-
2	-10.0	-C-C-C- : : : -G-G-G-
2	-10.0	-G-G-G- : : : -C-C-C-
1	-15.0	-C-C-C-C- : : : : -G-G-G-G-

Für das gleiche Beispiel produzierte HELIX.LISTE in der Zählversion die folgenden beiden Tabellen:

Helixzaehlung

Länge	Anzahl
2	5
3	13
4	1
5	1

Stabilität	Anzahl
0.0 >= Δ G >-2.0	7
-2.0 >= Δ G >-4.0	1
-4.0 >= Δ G >-6.0	7
-10.0 >= Δ G >-12.0	4
-14.0 >= Δ G >-16.0	1

fach, zweifach, dreifach usw. in der Bindungsmatrix auftreten:

Anzahl des Auftretens	Anzahl, der durch Basenpaarreihenfolge unterschiedenen Helices
1	89
2	21
3	9
4	3
5	6
6	2
7	1
8	1
9	2
10	0
11	1
12	2

Es gibt nur 3 verschiedene Helices, die mehr als 10 mal im Abschnitt 1-100 von MS2 auftreten.

Bis auf die Helix $\begin{matrix}-G-G-G-\\ \vdots\ \vdots\ \vdots\\ -U-U-U-\end{matrix}$ sind alle Helices, die häufiger als 5 mal auftreten kürzer als 3 Basenpaare. Es muss ferner erwähnt werden, dass Helices, die nur aus einem Basenpaar bestehen und solche, die nur aus (G·U)-Basenpaaren bestehen und kürzer als 3 Nucleotide sind aus später noch darzulegenden Gründen nicht in die Zählung mit aufgenommen wurden, weil nicht zur Stabilisierung einer Sekundärstruktur beitragen können und somit für weitere Untersuchungen, bzw. für das Aufstellen von Sekundärstrukturmodellen nicht verwendet werden können.

4.2.5 Auswertung der integralen Längenzählung nach LESK (HELIX.LISTE)

Das Programm HELIX.LISTE produziert eine Tabelle der Anzahlen aller möglichen helicalen Bereiche einer beliebig langen RNS, aufgeschlüsselt nach der Länge. Es ist nun möglich, diese Zahlen nach einem Verfahren von LESK (1974) (28) daraufhin zu untersuchen, ob sie von denjenigen, die nach Basenzusammensetzung und der Funktion pF der Faltungskapazität zu erwarten wären, signifikant abweichen. Daraus lassen sich dann Schlüsse ziehen, bezüglich der Anpassung des Moleküls an einen Zwang zu stabiler Sekundärstrukturbildung.

Bei der Ableitung des Erwartungswertes der integralen Anzahl der helicalen Bereiche mit der Länge n in einem Molekül der Gesamtlänge M halte ich mich an die Ausführungen von LESK (28). Dieser Erwartungswert wird mit $E_{int}(H^n)$ bezeichnet.
Hat die RNS die Länge M so sind genau M-n+1 verschiedene Sequenzen der Länge n aus ihr wählbar. Für das Beispiel der ersten 20 Nucleotide von MS2 wird M = 20 und es sind in dieser Sequenz genau 18 = 20 - 3 + 1 Teilsequenzen mit einer Länge von n = 3 Nucleotiden enthalten:

Sequenz:
-G-G-G-U-G-G-G-A-C-C-C-C-U-U-U-C-G-G-G-G-

Teilsequenzen à 3 Nucleotide:

1.	-G-G-G-	(1 bis 3)	10.	-C-C-C-	(10 bis 12)
2.	-G-G-U-	(2 bis 4)	11.	-C-C-U-	(11 bis 13)
3.	-G-U-G-	(3 bis 5)	12.	-C-U-U-	(12 bis 14)
4.	-U-G-G-	(4 bis 6)	13.	-U-U-U-	(13 bis 15)
5.	-G-G-G-	(5 bis 7)	14.	-U-U-C-	(14 bis 16)
6.	-G-G-A-	(6 bis 8)	15.	-U-C-G-	(15 bis 17)
7.	-G-A-C-	(7 bis 9)	16.	-C-G-G-	(16 bis 18)
8.	-A-C-C-	(8 bis 10)	17.	-G-G-G-	(17 bis 19)
9.	-C-C-C-	(9 bis 11)	18.	-G-G-G-	(18 bis 20)

Allein aus sterischen Gründen ist es nicht möglich, dass sich jede dieser Teilsequenzen mit jeder anderen zu einer Helix zusammenlagert, auch wenn die Basen der beiden Sequenzen zusammenpassen. Es muss erstens berücksichtigt werden, dass eine Sequenz nicht mit sich selbst in Kontakt treten kann und dass ebenfalls alle Sequenzen, die sich teilweise mit ihr überlappen für eine Helixbildung nicht in Frage kommen. Auch müssen die oben erwähnten Mindestabstände der Basen eines Basenpaares berücksichtigt werden. Hieraus folgt, dass nur solche Teilsequenzen miteinander eine Helix bilden können, die mindestens drei Nucleotide Zwischenraum besitzen. Diese drei Nucleotide bilden die kleinste mögliche Haarnadel (hair pin).

In der Bindungsmatrix wurden alle fortlaufenden Nucleotide einer RNS gegenübergestellt. Anstelle einer Gegenüberstellung von Nucleotiden kann man auch von einer Gegenüberstellung aller Teilsequenzen mit einer Länge von n = 1 sprechen. Analog wäre eine Gegenüberstellung aller fortlaufenden Teilsequenzen mit der Länge n = 2,3,4, usw. möglich. In Abb.23 ist die Bindungsmatrix der 18 oben aufgeführten 3er Sequenzen des Abschnittes 1-20 von MS2 wiedergegeben. Alle Kombinationen, die aus sterischen Gründen als unmöglich anzusehen sind, wurden schraffiert. Dort wo eine Basenpaarbindung zwischen zwei 3er Sequenzen möglich ist, wurde ein Kreis eingezeichnet, über die Art der Basenpaarbindung ob (A·U),(G·C) oder (G·U) ist damit nichts ausgesagt, denn bei der Kombination von 3er Sequenzen werden 3 Basenpaare gebildet, die meistens voneinander verschieden sind. Aus Abb.23 lässt sich entnehmen, dass die Anzahl von möglichen 3er Helices, d.h. die Summe aller möglichen Matrixelemente, die nicht schraffiert sind, gleich (20-2·3-3)·(20-2·3-2) ist. Für den allgemeinen Fall eines Moleküles der Länge M ist diese An-

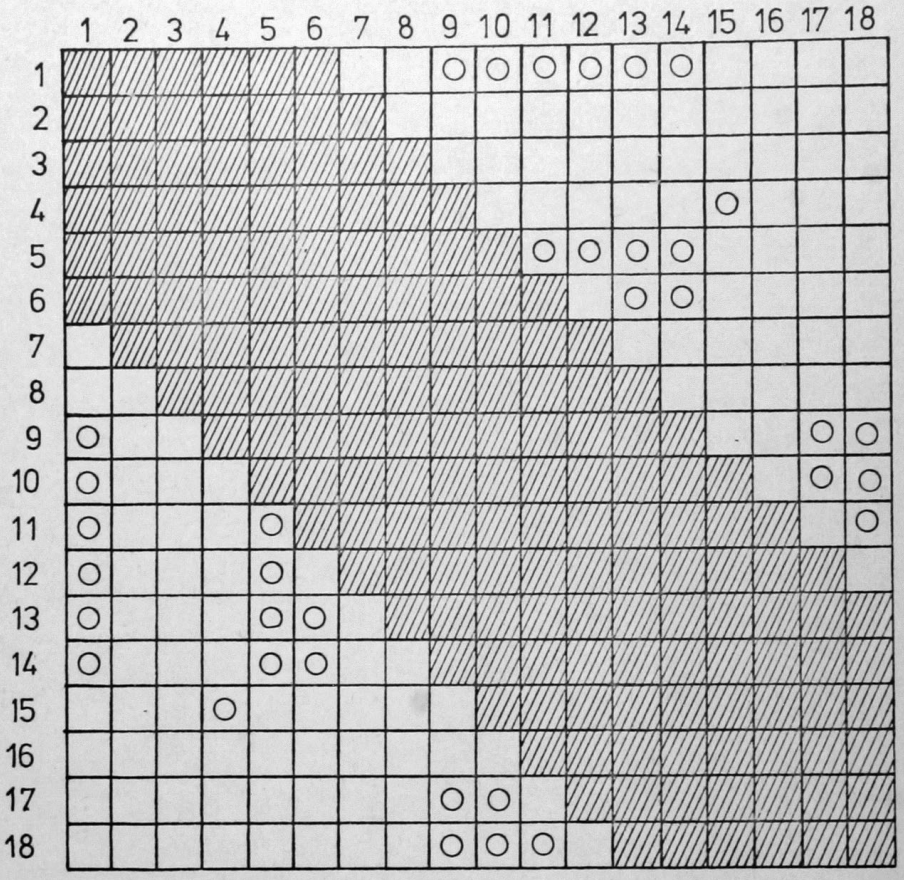


Abb.23: Bindungsmatrix aller möglichen 3er Sequenzen des Teilstückes 1-20 von MS2. Auflistung der 18 3er Sequenzen im Text.

zahl gleich $(M-2\cdot n-3)\cdot(M-2\cdot n-2)$. Diese Anzahl muss wegen der Symmetrie der Matrix in Bezug auf ihre Diagonale von links oben nach rechts unten noch halbiert werden, wie dies auch für die Bindungsmatrix der 1er Sequenzen (Abschnitt 4.2.1) geschehen ist. Die so berechnete Zahl wird in der Folge mit $K^n(M)$ bezeichnet (n ist die Länge der Teilsequenzen bzw. der Helices die sich aus ihnen bilden können und M ist die Länge der RNS).
Es ist somit:

$$K^3(20) = \frac{1}{2}\,(20-2\cdot 3-3)\cdot(20-2\cdot 3-2) = 66$$

und allgemein

$$K^n(M) = \frac{1}{2}\,(M-2\cdot n-3)\cdot(M-2\cdot n-2)$$

Aber nicht alle dieser Kombinationen von Sequenzen der Länge n ergeben Helices der Länge n, denn nur in einem Teil der Fälle sind die Sequenzen komplementär. Die Wahrscheinlichkeit, dass zwei beliebige Basen der RNS mineinander in Bindung treten können ist nach Abschnitt 4 gleich pF. Die Wahrscheinlichkeit der Bildung einer 2er Helix ist somit gleich pFpF oder pF^2, einer 3er Helix gleich pF^3 oder allgemein einer Helix der Länge n gleich pF^n. Von den möglichen Kombinationen wird also nach Erwartung nur der Anteil pF^n eine echte Komplementarität besitzen. Der Erwartungswert für die Anzahl der Helices der Länge n bei einer RNS der Länge M ist somit:

$$E_{int}(H^n) = \frac{1}{2}\cdot(M-2n-3)(M-2n-2)\cdot pF^n$$

"int"steht für "integralen Erwartungswert" im Gegensatz zu dem "absoluten Erwartungswert", $E_{abs}(H^n)$, der im nächsten Abschnitt definiert wird.

Es wurden nun für den Messenger MS2 alle vorkommenden Helices durch das Programm HELIX. LISTE gezählt und eine Liste der Anzahlen der Helices aller vorkommenden Längen angefertigt. Diese Zahlen können nun statistisch mit denjenigen Erwartungswerten verglichen werden, die sich mit der Formel für $E_{int}(H^n)$ für einen Messenger der Länge M=3569 (der Nucleotidanzahl von MS2) und der gleichen Basenzusammensetzung wie MS2 errechnen lassen. Für pF wird also der Wert aus Abschnitt 4.1.1 genommen um alle $E_{int}(H^n)$-Werte von MS2 für n = 1 bis 16 zu errechnen. Der pF-Wert von MS2 nach Basenzusammensetzung ist 0.3777. In der folgenden Tabelle sind die Zählwerte und Schätzwerte aller Helices der Längen 1 bis 16 aufgeführt. Die Schätzwerte wurden auf ganze Zahlen gerundet. Als dritte Spalte werden die Differenzen: Zählwerte minus Schätzwerte aufgeführt. Statis-

tisch gesehen entsprechen die Erwartungswerte $E_{int}(H^n)$ den Mittelwerten μ der Grundgesamtheiten und die Zählwerte den Mittelwerten $\bar{x}$ von Stichproben, die den Grundgesamtheiten entnommen wurden.

n	$\mu = E_{int}(H^n)$	$\bar{x}$ = Zählwerte	$\bar{x} - \mu$
1	2399702	⁄.	⁄.
2	905440	711249	⁄.
3	341635	343429	1794
4	128903	129489	586
5	47945	49213	1268
6	18351	18681	330
7	6924	7142	218
8	2613	2774	161
9	986	1062	80
10	372	402	30
11	140	140	0
12	53	50	-3
13	20	13	-7
14	8	5	-3
15	3	2	-1
16	1	0	-1

Die Anzahl der Helices mit der Länge 1, d.h. einzelne Basenpaare, wurden von HELIX.LISTE nicht gezählt, weiterhin wurde auch ein Grossteil der 2er Helices nicht gezählt, weil diese wie weiter unten gezeigt wird, für spätere Verwendung als Teil einer stabilen Gesamtsekundärstruktur des RNS-Moleküles nicht verwendbar sind. Die Werte für n=1 und n=2 werden bei der weiteren Auswertung nicht mehr berücksichtigt, die Häufigkeit der 2er Helices ist ja auch bereits weiter oben in den Abschnitten 4.1.5 und 4.1.6 statistisch untersucht worden, es könnte sich also hier nichts neues mehr ergeben. Es seien noch die Werte der

Funktion $K^n(3569)$ für n = 1 bis n = 16 aufgeführt, die als Stichprobenumfang in die statistische Analyse eingehen:

$K^1(3569) = 6352830$
$K^2(3569) = 6345703$
$K^3(3569) = 6338580$
$K^4(3569) = 6331461$
$K^5(3569) = 6324346$
$K^6(3569) = 6317235$
$K^7(3569) = 6310128$
$K^8(3569) = 6303025$
$K^9(3569) = 6295926$
$K^{10}(3569) = 6288831$
$K^{11}(3569) = 6281740$
$K^{12}(3569) = 6274653$
$K^{13}(3569) = 6267570$
$K^{14}(3569) = 6260491$
$K^{15}(3569) = 6253416$
$K^{16}(3569) = 6246345$

Diese Zahlen geben die Anzahlen der Kombinationen von je 2 Sequenzen von n Nucleotiden zu einer Helix in einer RNS der Länge 3569 wieder. Der statistische Test (u-Test) führt zu folgenden Werten für die Testgrösse u und das Integral der Normalverteilung von minus Unendlich bis zur Stelle u (= Pr(u)) :

n	u	Pr(u)
3	3.156	0.9992
4	1.649	0.9504
5	5.772	0.999999
6	2.440	0.9927
7	2.621	0.9956
8	3.151	0.9992
9	2.548	0.9946
10	1.556	0.9401
11	0.000	0.5000
12	-0.412	0.3402
13	-1.566	0.0587
14	-1.093	0.1372
15	-0.593	0.2766
16	-0.736	0.2309

Eigentlich wäre ein Absinken der Zählwerte gegenüber den Schätzwerten zu erwarten, da bei der Schätzung nur verlangt wird, dass die komplementären Sequenzen einen Mindestabstand von 3 Nucleotiden haben, in HELIX.LISTE aber bei Vorliegen einer (A·U) oder (G·U)-Bindung am rechten oberen Ende der Helix (in der Bindungsmatrix gesehen) der Mindestabstand der komplementären Sequenzen 4 Nucleotide beträgt. Es werden also in der tatsächlichen Zählung einige Helices kürzer sein als dies von der Schätzung her erwartet wird.
Trotzdem zeigt es sich, dass Steigerungen der Zählwerte gegenüber den Schätzwerten eintreten und in den Fällen n=3,5,6,7,8 und 9 als signifikant anzusehen sind.

Es ist nun aber so, dass die Schätzung der Helices nach LESK nicht die absoluten Anzahlen der möglichen Helices in einer RNS wiedergibt, sondern die "integralen" Anzahlen, d.h. eine Helix der Länge n wird als solche mit in den Erwartungswert aufgenommen, gleichgültig, ob sie isoliert in der Matrix steht oder Teil einer längeren zweiten Helix ist. So sind z.B. in einer Helix der Länge 4 zwei Helices der Länge 3"enthalten" oder drei Helices der Länge 2. Das heisst also anders gesagt, dass in dem Schätzwert $E_{int}(H^4)$ auch der Schätzwert $E_{int}(H^5)$ zweimal enthalten sein muss und der Schätzwert $E_{int}(H^6)$ dreimal usw.. Um dieser Tatsache gerecht zu werden, müssten die Original-Zählwerte des Programms HELIX.LISTE entsprechend umgerechnet werden. HELIX.LISTE produziert eine Liste des absoluten Auftretens aller Helices; will man hieraus das integrale Auftreten errechnen, muss man alle Vorkommen kürzerer Helices in längeren mitzählen. Es ergab sich z.B. dass die integrale Anzahl der 10er Helices in MS2 gleich

$$172 + 2\cdot53 + 3\cdot29 + 4\cdot5 + 5\cdot1 + 6\cdot2$$

ist, denn die absoluten Anzahlen der Helices

der Längen 11 bis 16 waren 53,29,5,1,2 und 0 und die 10er Helices kommen 2 mal in den 11ern und 3 mal in 12ern, 4 mal in 13ern,5 mal in den 14ern sowie 6 mal in den 15ern vor. Daraus ergibt sich, dass eine signifkante Steigerung der Häufigkeit der Helices der Länge n auch zu einer signifikanten Steigerung der Anzahl der Helices der Länge n-1 führen muss; Steigerungen akkumulieren sich also bei kürzeren n. Um zu einer statistischen Abschätzung der absoluten Anzahlen zu kommen musste die Formel für den Erwartungswert der absoluten Anzahlen gefunden werden (Abschnitt 4.2.6)

4.2.6 Auswertung der absoluten Längenzählung (HELIX.LISTE)

Die nun folgenden Ableitungen in den Kapiteln 4.2.6 und 4.2.7 gehen über die Angaben LESKs (28) hinaus.

Um den Erwartungswert für die absolute Anzahl von Helices der Länge n zu errechnen benötigt man die absolute Wahrscheinlichkeit $p_{abs}(H^n)$ für die Bildung einer Helix der Länge n. Zusammen mit der bereits bekannten Anzahl von Sequenzkombinationen aus zwei Segmenten der Länge n ($K^n(M)$) gibt dies den gewünschten Erwartungswert.

Jede RNS der Länge M kann nur Helices bis zu einer Maximallänge von n_{max} bilden. Es ist

$$n_{max} = \begin{cases} (M/2)-2 & \text{wenn M eine gerade Zahl ist} \\ ((M+1))/2)-2 & \text{wenn M eine ungerade Zahl ist} \end{cases}$$

Für Helices der Länge n_{max} ist die integrale Anzahl gleich der absoluten Anzahl, weil diese Helices in keinen noch längerkettigen Helices enthalten sein können. Die absolute Wahrscheinlichkeit $p_{abs}(H^{n_{max}-1})$ des Auftretens der Helices der Länge $n_{max}-1$ ergibt sich somit unter Berücksichtigung der Tatsache, dass $K^{n_{max}}(M)$ kleiner ist als $K^{n_{max}-1}(M)$ zu:

$$p_{abs}(H^{n_{max}-1}) = pF^{n_{max}-1} - 2\cdot(K^{n_{max}}(M)/K^{n_{max}-1}(M))\cdot p_{abs}(H^{n_{max}})$$

So lässt sich zu jeder Helix der Länge n, wenn die Werte $p_{abs}(H^{n+1})$ bis $p_{abs}(H^{n_{max}})$ bekannt sind, der Wert $p_{abs}(H^n)$ errechnen. Dies gibt uns die Möglichkeit, alle $p_{abs}(H^n)$ rekursiv zu berechnen, angefangen von dem bekannten $p_{abs}(H^{n_{max}})$ bis $p_{abs}(H^1)$.

Die dafür notwendige Rekursionsformel für Helices der Länge n ist:

$$p_{abs}(H^n) = pF^n - (K^{n+1}(M)/K^n(M))2p_{abs}(H^{n+1})$$
$$- (K^{n+2}(M)/K^n(M))3p_{abs}(H^{n+2})$$
$$- (K^{n+3}(M)/K^n(M))4p_{abs}(H^{n+3})$$
$$- \ . \ . \ .$$
$$- (K^{n_{max}}(M)/K^n(M))(n_{max}-n+1)p_{abs}(H^{n_{max}}) \ .$$

oder

$$p_{abs}(H^n) = pF^n - \sum_{f=n+1}^{n_{max}} (K^f(M)/K^n(M))(f-n+1)p_{abs}(H^f) \ .$$

Wenn wir mit $E_{abs}(H^n)$ den Erwartungswert der absoluten Anzahl der Helices der Länge n bezeichnen, der gleich $p_{abs}(H^n) \cdot K^n(M)$ ist, so erhält man aus der obigen Formel folgende Umformung:

$$E_{abs}(H^n) = pF^n \cdot K^n(M) - \sum_{f=n+1}^{n_{max}} (f-n+1) \cdot E_{abs}(H^f)$$

oder mit

$$E_{int}(H^n) = pF^n \cdot K^n(M)$$

ergibt sich

$$E_{abs}(H^n) = E_{int}(H^n) - \sum_{f=n+1}^{n_{max}} (f-n+1) \cdot E_{abs}(H^f)$$

Mit Hilfe dieser Rekursionsformel gelang es sämtliche Werte von $p_{abs}(H^1)$ bis $p_{abs}(H^{n_{max}})$ zu rekonstruieren. n_{max} hat bei MS2 den Wert 1783, dies ist die längste kontinuierliche Helix, die

sich in MS2 bilden könnte.
Es seien zunächst die Werte $p_{abs}(H^n)$ für n = 1 bis n = 16 aufgeführt:

n	$p_{abs}(H^n)$
1	0.1464629173
2	0.0553247891
3	0.0208981521
4	0.0078940205
5	0.0029818856
6	0.0011263618
7	0.0004254710
8	0.0001607170
9	0.0000607085
10	0.0000229319
11	0.0000086623
12	0.0000032720
13	0.0000012360
14	0.0000004669
15	0.0000001764
16	0.0000000666

Hieraus konnten die Erwartungswerte $E_{abs}(H^n)$ ermittelt werden. Die folgende Tabelle gibt diese Erwartungswerte, die wiederum den Mittelwerten von statistischen Grundgesamtheiten entsprechen und die ursprünglichen Zählwerte nach HELIX.LISTE, die man Mittelwerten von Stichproben gleichsetzen kann, und die jeweiligen Differenzen:

n	$\mu = E_{abs}(H^n)$	$\bar{x}$ = absolute Zählwerte	$\bar{x} - \mu$
1	930454.38	⁄.	⁄.
2	351074.69	153880	⁄.
3	132464.63	133664	1199.37
4	49980.70	49744	-236.70
5	18858.48	18993	134.52
6	7115.49	7171	55.51
7	2684.78	2656	-28.78

n	$\mu = E_{abs}(H^n)$	$\overline{x}$ = absolute Zählwerte	$\overline{x} - \mu$
8	1013.00	1052	39.00
9	382.21	398	15.78
10	144.21	172	27.79
11	54.41	53	-1.41
12	20.53	29	8.47
13	7.75	5	-2.75
14	2.92	1	-1.92
15	1.10	2	0.90
16	0.42	0	-0.42

In wieweit sind die Unterschiede zwischen den Erwartungswerten und den tatsächlichen Zählwerten nach HELIX.LISTE im Falle MS2 statistisch signifikant? Es wurde wiederum ein u-Test durchgeführt:

n	u	Pr(u)
3	3.3304	0.9996
4	-1.0630	0.1439
5	0.9811	0.8367
6	0.6584	0.7449
7	-0.5558	0.2892
8	1.2253	0.8898
9	0.8072	0.7902
10	2.3190	0.9898
11	-0.1911	0.4242
12	1.8698	0.9692
13	-0.9892	0.1613
14	-1.1228	0.1308
15	0.8571	0.8043
16	-0.6563	0.2558

Es zeigt sich, dass nur die Steigerungen bei n=3 und n=10 bei unserer allgemein für diese Arbeit akzeptierten Signifikanzgrenze von 5% als statistisch abweichend beurteilt werden können. Alle anderen Werte liegen innerhalb der 95% aller Werte in der Nähe des Mittelwertes.

4.2.7 Versuch der erneuten Berechnung der Informationskapazität

Die Informationskapazität wurde bereits auf zweierlei verschiedene Arten berechnet, zum einen wurde die RNS als einfacher Informationsträger aufgefasst, bei der jeder Buchstabe die gleiche Wahrscheinlichkeit hat, an jeder Stelle des Trägers zu stehen, d.h. man berücksichtigte nur die Basenzusammensetzung des Informationsträgers und setzte den Prozentsatz eines Buchstabens, einer Base, gleich seiner Wahrscheinlichkeit an eine bestimmte Stelle der RNS zu geraten. Zum Andern fassten wir die RNS als MARKOFF-Kette auf, wo die Wahrscheinlichkeit eines Buchstabens, an eine bestimmte Stelle zu kommen, von dem vorhergehenden Buchstaben abhing, also ein Einfluss der Nachbarbasen vorlag. In beiden Fällen konnte anhand von Formeln der Informationstheorie ein mittlerer Informationsgehalt pro Base festgestellt werden und für unsere beiden untersuchten Beispiele, MS2 und ϕX174 konnte nur eine sehr geringe Verringerung der Codierungsfähigkeit der RNS, bzw. ihres Informationsgehaltes, festgestellt werden. Die mittlere Informationskapazität pro Base sank kaum unter den theoretischen Maximalwert von 2.000 bit.
Nach den Betrachtungen der letzten Kapitel zeichnet sich die RNS der Bakteriophagen aber nicht nur durch eine erhöhte Faltungskapazität (pF) oder einen erhöhten Faltungsgrad der 2er Sequenzen aus, sondern auch durch teilweise erhöhtes Vorkommen von Helices bestimmter Längen. Es fragt sich nun, wie man dieses Phänomen einer informationstheoretischen Untersuchung unterziehen kann. In diesem Kapitel wird versucht, die Codierfähigkeit oder Brauchbarkeit als Informationsträger der RNS dadurch zu untersuchen, indem festgestellt wird, wieviel verschiedene Sequenzen eine RNS noch haben kann, wenn sie einem evolutionären Druck in Richtung auf bestimmte Absolut-Häufig-

keiten bestimmter Helixsorten nachgeben muss. Oder anders gefragt, wie hoch ist die Wahrscheinlichkeit, dass eine beliebige Sequenz mindestens k Helices der Länge n bilden kann? Um diese Frage zu beantworten werden zunächst die Fälle k = 1 und k = 2 betrachtet und dann auf den allgemeinen Fall geschlossen.

1. k = 1

Wie hoch ist die Wahrscheinlichkeit, dass in einer beliebigen Sequenz mindestens eine Helix der Länge n enthalten ist?
In einer RNS gibt es $K^n(M)$ verschiedene mögliche Kombinationen von Sequenzen der Länge n. Es können somit an $K^n(M)$ verschiedenen Stellen im Molekül Helices der Länge n gebildet werden. Diese Stellen seien durchnummeriert und bekommen den allgemeinen Index i so dass gilt:

$1 <= i <= K^n(M)$.

Dann sei mit H_i die Helix an der i-ten Stelle gemeint. Die Anzahl aller möglichen verschiedenen Sequenzen beträgt 4^M, weil an jeder Stelle des Moleküls 4 verschiedene Basen stehen können. Tritt aber an der Stelle i eine Helix der Länge n auf, so gibt es im Rest des Moleküls nur noch 4^{M-2n} verschiedene Sequenzen, an der Stelle i (d.h. innerhalb der beiden Teilsequenzen, die die Helix der Länge n aufbauen) können 6^n verschiedene Helices auftreten, weil es an jeder Stelle der Helix 6 verschiedene Basenpaare geben kann ((G·U),(U·G),(A·U), (U·A),(G·C) und (C·G)). Die Gesamtzahl aller Sequenzen, bei denen an der Stelle i eine Helix der Länge n auftritt, ist somit:

$4^{M-2n} \cdot 6^n$

Die Wahrscheinlichkeit ($p_{int}(H_i^n)$), dass in einer beliebigen Sequenz die Helix H_i^n auftritt, ist also:

$$p_{int}(H_i^n) = \frac{4^{M-2n} \cdot 6^n}{4^M} = 4^{-2n} \cdot 6^n \approx 4^{-0.7075n}$$

Hieraus folgert zunächst, dass $p_{int}(H_i^n)$ von der Grösse M unabhängig ist. Gesucht ist allerdings nicht die Wahrscheinlichkeit, dass sich eine bestimmte Helix i bildet, sondern nur die Wahrscheinlichkeit $p_{int}(H^n)$, dass sich irgendeine der $K^n(M)$ verschiedenen möglichen Helices bildet. Es ist also:

$$p_{int}(H^n) = p_{int}(H_1^n \vee H_2^n \vee \ldots \vee H_i^n \vee \ldots \vee H_{K^n(M)}^n)$$

bzw.

$$p_{int}(H^n) = p_{int}\left(\bigvee_{i=1}^{K^n(M)} H_i^n\right)$$

Fernerhin gilt:

$$p_{int}(H^n) = 1 - p_{int}(\neg H^n)$$

und

$$p_{int}(\neg H^n) = p_{int}\left(\neg\left(\bigvee_i H_i^n\right)\right) = p_{int}\left(\bigwedge_i (\neg H_i^n)\right)$$

Nach den Regeln der Wahrscheinlichkeitsrechnung gilt für zwei u n a b h ä n g i g e Ereignisse A und B:

$p(A \wedge B) = p(A) \cdot p(B)$

So gilt auch in unserem Fall:

$$p_{int}(H^n) = 1 - (p_{int}(\neg H_1^n) \cdot p_{int}(\neg H_2^n) \cdot \ldots \cdot p_{int}(\neg H_{K^n(M)}^n))$$

bzw.

$$p_{int}(H^n) = 1 - \prod_i p_{int}(\neg H_i^n)$$

mit

$p_{int}(\neg H_i^n) = 1 - p_{int}(H_i^n)$ ergibt sich:

$$p_{int}(H) = 1 - \prod_i (1 - p_{int}(H_i^n)) \approx 1 - \prod_i (1 - 4^{-0.7075n})$$

$$= 1 - (1 - 4^{-0.7075n})^{K^n(M)}$$

Letzteres ist die gesuchte Wahrscheinlichkeit, dass sich mindestens eine Helix der Länge n in

einer beliebigen Sequenz der Länge M bildet.

2. $k = 2$

Wie hoch ist die Wahrscheinlichkeit, dass in einer beliebigen Sequenz mindestens zwei Helices der Länge n enthalten sind? Aus der Menge der möglichen Helices mit n Basenpaaren seien die Helices H_i^n und H_j^n herausgegriffen. Für j gilt wie für i: $1 <= j <= K^n(M)$

Wenn man die 4n Nucleotide, die zur Bildung dieser beiden Helices notwendig sind, aus einem Molekül herausnimmt, bleiben für die restlichen M-4n Nucleotide noch 4^{M-4n} verschiedene Sequenzen. Die beiden Helices können jede 6^n verschiedene Basenpaarzusammensetzungen haben, zusammen sind dies wiederum 6^{2n} verschiedene Molekülzustände, also gibt es insgesamt

$$4^{M-4n} \cdot 6^{2n}$$

verschiedene Sequenzen mit den beiden Helices H_i^n und H_j^n.

Die Wahrscheinlichkeit, dass eine beliebige Sequenz die Helices H_i^n und H_j^n enthält ist:

$$p_{int}(H_i^n \wedge H_j^n) = \frac{4^{M-4n} \cdot 6^{2n}}{4^M} = 4^{-4n} \cdot 6^{2n} \approx 4^{-1.4150n}$$

Gesucht ist die Mindestwahrscheinlichkeit, dass zwei Helices der Länge n in der Sequenz verankert sind:

$$p_{int}(2H^n) = p_{int}((H_1^n \wedge H_2^n) \vee (H_1^n \wedge H_3^n) \vee \ldots$$

$$\ldots \vee (H_{K^n(M)}^n \wedge H_{K^n(M)-1}^n))$$

bzw.

$$p_{int}(2H^n) = p_{int}\left(\bigvee_{i,j} (H_i^n \wedge H_j^n) \right) \qquad \text{mit } i \neq j$$

Die letzte Summation erfolgt über alle i und j, die nicht einander gleich sind, dies würde bedeuten, dass dieselbe Helix zweimal in einer Sequenz auftreten sollte. Weiterhin ist

$$p_{int}(2H^n) = 1 - p_{int}(\neg 2H^n)$$

$$= 1 - p_{int}(\neg(\bigvee_{i,j}(H_i^n \wedge H_j^n))$$

$$= 1 - p_{int}(\bigwedge_{i,j}(\neg(H_i^n \wedge H_j^n))$$

$$= 1 - \prod_{i,j} p_{int}(\neg(H_i^n \wedge H_j^n))$$

$$= 1 - \prod_{i,j}(1 - p_{int}(H_i^n \wedge H_j^n))$$

Die Anzahl der Kombinationen von i und j mit der Bedingung, dass i nicht gleich j sein darf, entspricht der Anzahl von möglichen Auswahlen von 2 Elementen aus einer Grundmenge von $K^n(M)$ Elementen ohne Wiederholung eines Elementes und ohne Berücksichtigung der Reihenfolge der herausgegriffenen Elemente. Diese Anzahl wird nach folgender Formel aus der Kombinatorik berechnet:

$$\binom{K^n(M)}{2} = \frac{K^n(M)!}{2! \cdot (K^n(M)-2)!}$$

.

Der zweite Summand in der Formel für $p_{int}(2H^n)$ besitzt also $\binom{K^n(M)}{2}$ verschiedene Faktoren.

$$p_{int}(2H^n) = 1 - (1 - p_{int}(H_i^n \wedge H_j^n))^{\binom{K^n(M)}{2}}$$

$$\approx 1 - (1 - 4^{-1.4150n})^{\binom{K^n(M)}{2}}$$

Dies ist die gesuchte Formel für die Mindestwahrscheinlichkeit von 2 Helices mit je n Basenpaaren.

3. Allgemeiner Fall für k

Wie gross ist die Wahrscheinlichkeit, dass in einer beliebigen Sequenz mindestens k Helices der Länge n enthalten sind? Entsprechend der Untersuchung für die Fälle k = 1 und k = 2 ist die Wahrscheinlichkeit, dass in einer Sequenz genau k herausgewählte Helices H_i^n vorhanden sind, gleich

$$p_{int}\left(\bigwedge_{g=1}^{k} H_{i_g}^{n}\right) = \frac{4^{M-2kn} \cdot 6^{kn}}{4^{M}} \approx 4^{-0.7075kn}$$

und die Mindestwahrscheinlichkeit, dass überhaupt k Helices in einer Sequenz enthalten sind, ist

$$p_{int}(kH^n) = 1 - \left(1 - p_{int}\left(\bigwedge_{g=1}^{k} H_{i_g}^{n}\right)\right)^{\binom{K^n(M)}{k}}$$

$$\approx 1 - \left(1 - 4^{-0.7075kn}\right)^{\binom{K^n(M)}{k}}$$

mit

$$\binom{K^n(M)}{k} = \frac{K^n(M)!}{k! \cdot (K^n(M)-k)!}$$

Bei dem statistischen Test der absoluten Anzahlen aller Helices mit bestimmten Längen ergab sich dass bei MS2 die Helices mit den Längen 3 und 10 signifikant erhöht sind. Um die Frage beantworten zu können, ob dadurch die Informationskapazität verringert ist, interessiert die Ausrechnung der Formel für $p_{int}(kH^n)$ für die Fälle n=3 und n=10.

Mit k ist die integrale Anzahl an Helices gemeint, man kann daher für k nicht den Wert 172 einsetzen, der z.B. die signifikant gestiegene absolute Anzahl der 10er Helices darstellte, sonder den Wert 756, der um die Vorkommen von 10er Helices in den Helices grösser als 10 korrigiert wurde. Die dafür zugrundeliegenden Anzahlen der längerkettigen Helices wurden der Tabelle der Erwartungswerte der absoluten Anzahlen entnommen, damit der statistische Test der 10er Helices nicht von gestiegenen oder gesunkenen Anzahlen längerkettiger Helices beeinflusst wird. Auf die selbe Weise wird die Anzahl 671332 für die 3er Helices errechnet, auch hier sind die erwartungsgemässen Vorkommen in längerkettigen Helices zur absoluten Anzahl hinzugezählt worden.

Es wird sodann:

$$p_{int}(671332H^3) \approx 1-(1 - 4^{-1424939.93})^{\binom{6338580}{671332}}$$

und

$$p_{int}(\quad 756H^{10}) \approx 1-(1 - 4^{-\quad 5348.86})^{\binom{6288831}{756}}$$

Die Berechnung dieser Formeln stellt ein beträchtliches numerisches Problem dar. Mit Hilfe der STIRLINGschen Formel für Fakultäten ergaben sich folgende Schätzwerte:

$$\binom{6338580}{671332} \approx 1.408 \cdot 10^{930122}$$

$$\binom{6288831}{756} \approx 1.058 \cdot 10^{3290}$$

Bei der weiteren Berechnung bedarf es einiger Umformungen. Wenn

$x = 4^{-0.7075kn}$ und $y = \binom{K^n(M)}{k}$ gesetzt wird,

ergibt sich

$$p_{int}(kH^n) = 1 - (1 - x)^y$$

Für die Funktion $\ln(1 - x)$ gibt es im Bereich $-1 <= X < 1$ die konvergente Reihenentwicklung:

$$\ln(1 - x) = - x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots$$

Eingesetzt in die obige Formel führt dies zu:

$$p_{int}(kH^n) = 1-e^{y\cdot\left(- x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots\right)}$$

oder

$$p_{int}(kH^n) = 1-e^{-yx - \frac{yx^2}{2} - \frac{yx^3}{3} - \ldots}$$

yx ist im Fall von n=3 ungefähr gleich 10^{72222} und im Fall von n=10 ungefähr 10^{70}. In beiden Fällen ist also keine nennenswerte Verringerung der Anzahl der möglichen Sequenzen zu verzeichnen. Dies ist im Gegensatz zu dem Ergebnis von BALL (1972)

(2) der zu sehr einschneidenden Verringerungen dieser Anzahl gelangte. BALL stellte bei R17, einem mit MS2 nah verwandten Bakteriophagen mit fast derselben Sequenz wie MS2, fest, dass, bezogen auf die Menge aller möglichen Sequenzen dieses Phagen, nur eine Sequenz aus 10^{238} das gleiche Ausmass an Basenpaarbindung (73%) aufweist, wie R17 selbst. In dieser Untersuchung wird aber angenommen, dass sich die helicalen Bereiche an bestimmten Stellen bilden müssen, es wird nicht über alle $K^n(M)$ möglichen Stellen summiert; diese Annahme führt zu bedeutend stärkeren Selektionsdrücken als die in diesem Kapitel verwendete.

In dieser Arbeit wird angenommen, dass irgendeine Sekundärstruktur vorliegen muss, BALL nimmt aber an dass eine bestimmte Sekundärstruktur vorliegen muss.

4.2.8 Auswertung der Stabilitätszählung (HELIX.LISTE)

Gleichzeitig mit der Auszählung der Helices nach ihrer Länge fertigt HELIX.LISTE eine Tabelle mit den Anzahlen der Helices verschiedener Stabilität an. Die Stabilitäten wurden nach den Angaben von TINOCO et al.(38) über die freien Energien der Bildung bestimmter Basenpaarsequenzen, berechnet. Diese Energien sind in kcal, wurden aber hier in kJ umgerechnet (Deshalb die "krummen" Zahlen!). Folgendes ist eine Liste der Anzahl von Helices mit bestimmter freier Energie von MS2:

freie Energie(kJ)	Anzahl
0.000 <=ΔG< -8.371	98846
-8.371 <=ΔG< -16.742	158368
-16.742 <=ΔG< -25.113	72395
-25.113 <=ΔG< -33.484	17893
-33.484 <=ΔG< -41.855	9965
-41.855 <=ΔG< -50.226	5924
-50.226 <=ΔG< -58.597	2742
-58.597 <=ΔG< -66.968	1015
-66.968 <=ΔG< -75.339	306
-75.339 <=ΔG< -83.710	223
-83.710 <=ΔG< -92.081	68
-92.081 <=ΔG<-100.452	42
-100.452 <=ΔG<-108.823	22
-108.823 <=ΔG<-117.194	2
-117.194 <=ΔG<-125.565	6
-125.565 <=ΔG<-133.936	2
-133.936 <=ΔG<-142.307	1

Die Helices mit einer freien Energie grösser als -8.371 kJ sind, wie bereits bemerkt, unterrepresäntiert, weil sie bei der später zu erfolgenden Konstruktion einer Faltung der RNS von MS2 keine Verwendung finden können.
Es gibt keinen direkten Zusammenhang zwischen der Länge einer Helix und ihrer freien Energie.

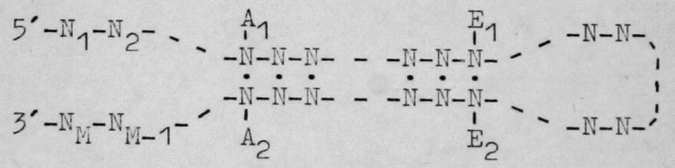
Sind in einer Helix genügend (G·U)-Basenpaare eingefügt, kann sie beliebig lang sein, ohne dass ihre Stabilität nach unseren Schätzwerten nennenswert wächst. Andererseits kann eine 3er Helix bereits eine freie Energie von -41.855 kJ haben, wenn sie nur aus (G·C)-Paaren besteht. Aus denselben Gründen kann aber auch eine Helix mit einer Stabilität von -125.565 kJ nicht kürzer als 7 Basenpaare sein.

4.3 Einfache Optimierung der Sekundärstruktur von RNS-Molekülen

Es wurde nach dem Prinzip der Maximumwahl ein Algorithmus geschaffen, der eine hypothetische Sekundärstruktur für ein beliebiges RNS-Molekül erstellt. Der Algorithmus wählt aus der Menge aller möglichen Helices eine Helix aus, schliesst sodann sterisch unmöglich gewordene Helices von der weiteren Betrachtung aus und wählt erneut unter den verbliebenen Helices die maximal stabile aus. Dieser Vorgang wird solange wiederholt, bis keine weiteren Helices mehr verblieben sind. Die herausgewählten Helices werden in den Sekundärstrukturvorschlag einbezogen. Diesen Vorgang im Programm habe ich "falten" genannt, eine herausgewählte Helix wird "gefaltet". Nimmt das Programm eine Helix aus dem Sekundärstrukturvorschlag wieder heraus, so wird sie "entfaltet".

4.3.1 Prinzipien der Faltung von RNS-Molekülen, dargestellt anhand der Bindungsmatrix

Zum Bilden einer Helix wird ein Nucleotidstrang der Form $5'\text{-}N_1\text{-}(N\text{-})_{M-2}N\text{-}3'$ in sich selbst zurückgefaltet, so dass sich eine intramolekulare Doppelhelixkonformation ergibt:



Vereinbarungsgemäss wird in dieser Arbeit das erste (vom 5'-Ende aus gesehen) in der Helix gebundene Nucleotid mit der Nummer A_1, das letzte entsprechend mit der Nummer A_2 (A wie Anfang). Diese beiden Nucleotide stellen das l i n k e Ende der Helix dar. Die beiden Nucleotide des r e c h t e n Endes haben, wie der

obigen Zeichnung zu entnehmen ist, die Nummern E_1 und E_2 (E wie Ende). Es gilt also: $A_1 < E_1 < E_2 < A_2$. Da eine Helix die Mindestlänge 2 hat, muss somit A_1 ungleich E_1 und E_2 ungleich A_2 sein. Aus sterischen Gründen gilt: $E_1+3 < E_2$. Definitionsgemäss sind die Nucleotide mit den Nummern E_1+1 bis E_2-1 die zu dieser Helix entsprechende Haarnadelschleife(hairpin-loop). In der Bindungmatrix lässt sich die Situation auf folgende Weise darstellen:(Abb.24):

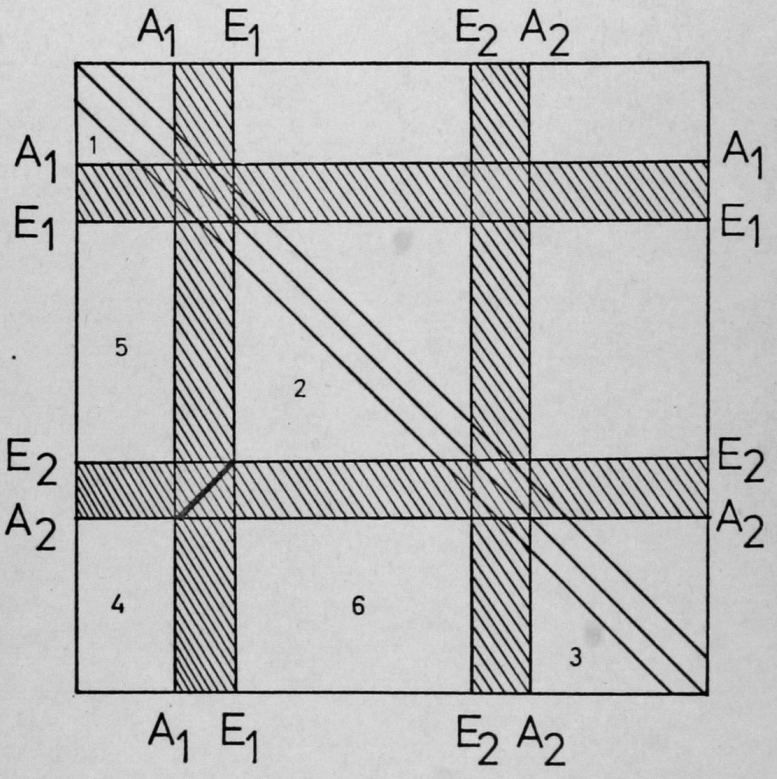


Abb. 24:Durch Aufnahme einer Helix (breite schwarze Linie) in einen Faltungsvorschlag können alle Basenpaare, die in den schraffierten Bereichen liegen, nicht mehr zusätzlich aufgenommen werden. Definition der Koordinaten A_1,E_1,E_2 und A_2 siehe Text.

Wenn man eine bestimmte Helix in ein Molekül inkorporiert, dann heisst das gleichzeitig, dass alle anderen möglichen Bindungen der Nucleotiede A_1 bis E_1 und E_2 bis A_2 ausgeschlossen werden. Die in der Abb.24 schraffierten Flächen beinhalten also diejenigen Basenpaare, die in eine derartig "vorgefaltete" RNS nicht

mehr inkorporiert werden dürfen. Die verbleibenden, aufgrund der Basenidentität noch möglichen Bindungen liegen in 6 verschiedenen Bebieten, die von 1 bis 6 durchnumeriert sind. Die im Bereich 1 verbliebenen Helices oder Helixteile sind beschränkt auf den Anfang des Moleküles, nämlich die Nucleotide 1 bis A_1-1, entsprechend sind die Helices aus Bereich 3 "lokal" auf die Nucleotide A_2+1 bis M, dem Ende des Moleküls, beschränkt. Der Bereich 2 umfasst diejenigen Helices, die sich lokal zur Haarnadelschleife bilden und Bereich 4 stellt Kombinationen der beiden Enden des Moleküls dar. In Bereich 5 und 6 werden jedoch entweder Nucleotide aus dem ungebundenen Anfang mit Nucleotiden innerhalb der Haarnadelschleife kombiniert, oder Nucleotide des ungebundenen Endes mit solchen der Haarnadelschleife. Faltungen, bei denen Helices aus den Bereichen 5 oder 6 in einem Molekül inkorporiert waren, sind bereits veröffentlicht worden, allerdings nur im Rahmen einer theoretischen Arbeit (JORDAN,(1971) (21)). Experimentell hat sich eine solche Struktur noch nicht feststellen lassen. Beim Vorschlag derartiger Konformationen als mögliche Sekundärstrukturen für tRNS-Moleküle ist nicht überprüft worden, ob aufgrund der tatsächlichen physikalischen Abstände zwischen den Basen und den Begrenzungen, denen eine Verzerrung des Moleküls unterliegt, eine Bindung der genannten Art überhaupt auftreten kann. Bei Bindungen aus den Bereichen 1 bis 4 braucht eine solche Überprüfung nicht unbedingt durchgeführt zu werden weil hier kaum sterische Probleme auftreten. Weiterhin müsste bei einer Konformation mit Helices aus den Bereichen 5 und 6 gesichert sein, dass sich die beiden in Kontakt tretenden Teile des Moleküls gegenseitig spiralig umwickeln können, wie dies die korrekte Bildung einer Doppelhelix erfordern wurde. Die Überprüfung, ob in einem gegebenen Fall der Einbau einer Helix aus den Bereichen

5 und 6 in ein Molekül überhaupt physikalisch möglich ist, muss anhand eines Modells geführt werden, in das all die genannten physikalischen Parameter eingehen, ein Modell also, dass viel mehr einer realen RNS entspräche, als die einfache Kette der Nucleotidbuchstaben, die die Grundlage für alle Betrachtungen im Rahmen dieser Arbeit darstellt. Ein solches Modell stellt computertechnisch gesehen einen enorm grösseren Aufwand dar, als die reine Untersuchung der möglichen Basenpaarkombinationen, die hier versucht worden ist. Es wird also im Rahmen der Überlegungen dieser Arbeit willkürlich vereinbart, dass sich Basenpaarkombinationen aus den Bereichen 5 und 6 nicht bilden können, da solche Kombinaionen hier nicht auf ihre physikalische Realisierbarkeit untersucht werden können und aller Voraussicht nach nur sehr selten, wenn überhaupt, auftreten.

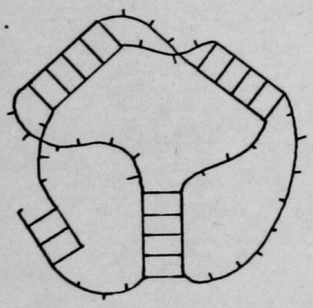


Abb.25: Faltung mit Basenpaarkombinationen aus den Bereichen 5 und 6 von Abb. 24, die zwar auf dem Papier leicht darzustellen ist, deren physikalische Realisierbarkeit aber nur schwer überprüfbar ist.

Alle Sekundärstrukturen, die keine Basenpaare aus den Bereichen 5 und 6 enthalten, haben die Eigenschaft, dass man sie ohne Schwierigkeiten in der 2-dimensionalen Ebene des Zeichenblattes darstellen kann, ohne dass man gezwungen ist, Teile des Moleküles sich überkreuzen zu lassen. Deshalb werden diese Faltungen oder Sekunstrukturen als p l a n a r bezeichnet. Die von der Arbeitsgruppe FIERS vorgeschlagene Sekundärstruktur für den Bakteriophagen MS2 ist planar, wie man der Abbildung, die im hinteren Umschlagblatt eingefügt ist, entnehmen kann.

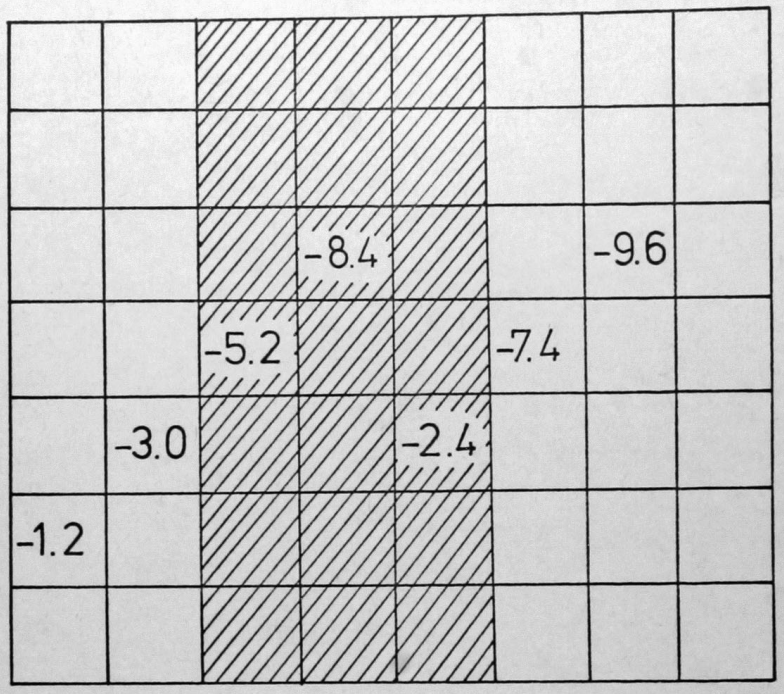
Wenn man versucht Sekundärstrukturen zu bilden, indem man eine bestimmte Helix aus der Menge der möglichen herausgreift, dann muss sichergestellt sein, dass die nächste Helix nicht aus den "verbotenen" Zonen der Matrix kommt, also zum einen nicht aus den schraffierten Flächen der Abb. 24 und zum anderen nicht aus den vereinbarungsgemäss nicht zugelassenen Flächen mit den Nummern 5 und 6.

4.3.2 Beschreibung der Computerverfahren zur Auswahl einer Helix und darauffolgendem Ausschluss sterisch unmöglich gewordener Helices

Durch die unterschiedliche Repräsentation der Menge der möglichen Helices im FORTRAN Programm FALTUNG und den SIMULA Programmen HELIX. LISTE und SIM.FALTUNG ergibt sich auch eine unterschiedliche Methodik der Auswahl der geeigneten Helices und Begrenzung der weiterhin möglichen helicalen Bereiche.

4.3.2.1 Auswahl und Begrenzung der Helices im Programm FALTUNG

In FALTUNG gibt es zwei M mal M Matrices; in der ersten, Matrix B, werden die Stapelkräfte aller möglichen 2er Helices gespeichert, wie dies in Abschnitt 4.2.3.1 beschrieben worden ist, in der zweiten, MatriX C, werden die von links unten nach rechts oben aufsummierten Stapelkräfte gespeichert. Die Gesamt-freie Energie einer Helix ist an der Matrixstelle E_1,E_2 d.h. am rechten Ende abgespeichert. Das Programm sucht nun nach dem numerisch niedrigsten Wert in der Matrix, dies ist gleichzeitig die Gesamt-freie Energie der stabilsten Helix der betreffenden RNS. Als nächstes werden die Stapelkräfte der sterisch nicht mehr möglichen 2er Helices in der Matrix B gelöscht und die Matrix C erneut berechnet. Es ist nicht möglich, einfach die entsprechenden Werte in C zu löschen, weil Helices, die mit einem Ende in einen sterisch "verbotenen" Bereich kommen, nur im Fall, dass dieses Ende das rechte war, bei einer Löschung noch auf dem Platz E_1,E_2 die richtige Gesamtstabilität aufweisen. In der Abb.26 wäre nach Löschung des schraffierten Streifens in der linken Helix der Wert -3.0 kcal die richtige Gesamtstabilität der verkürzten Helix, im rechten Fall wäre -9.6 unzutreffend, richtig aber -2.2 kcal.



Äbb.26: Ausschnitt aus der Matrix C der Programms FALTUNG, der den Effekt einer möglichen Löschung von sterisch unmöglich gewordenen Elementen auf die aufsummierten Stapelkräfte zeigt. C muss nach jeder Löschung in der Matrix B neu berechnet werden.

Der Vorgang der Helixauswahl, Löschung in B und Neuberechnung von C wiederholt sich solange, bis in C nur noch der Grundwert 0.0 vorhanden ist. Durch den Zwang, gesamte Bindungsmatrices zu bearbeiten und stetig neu zu berechnen, wird diese Methode nicht nur viel Speicherplatz, sonder auch viel Rechenzeit benötigen, so dass hiermit auf ökonomische Weise nur RNS-Moleküle bis zu einer Maximallänge von 200 Nucleotiden bearbeitet werden können. Die Durchschnittslänge einer tRNS ist 77 Nucleotide, fällt also in diesen Bereich, der Bakteriophage MS2 mit einer Länge von 3569 Nucleotiden kann hiermit nicht oder höchstens abschnittsweise gefaltet werden.

4.3.2.2 Auswahl und Begrenzung der Helices im Programm SIM.FALTUNG

Das Programm HELIX.LISTE kann als output die Menge der möglichen Helices für jede beliebige RNS in Form von Zahlentripeln produzieren. Diese Zahlentripel bestehen aus A_1,A_2 und der Länge der Helix. Diese Liste von Tripeln dient als input für das SIMULA-Programm SIM. FALTUNG, das sich hieraus die thermodynamisch stabilste Helix heraussucht. Besteht die Möglichkeit,zwischen verschiedenen gleich stabilen Helices zu wählen, so wird die längste Helix erwählt, bestehen auch hier wieder Mehrdeutigkeiten, so wird unter Zuhilfenahme eines Zufallszahlengenerators eine beliebige Helix unter der verbleibenden Gruppe der stabilsten und maximal langen Helices ausgewählt.
Nach der Auswahl der optimalen Helix werden die Zahlentripel, die programmintern als 2-dimensionaler Vektor dargestellt sind, auf Kompatibilität zur erwählten Helix hin untersucht. Ist eine Helix inkompatibel, so werden ihre Werte auf Null gesetzt, ist sie kompatibel, so bleiben die Werte unverändert. Ist sie partiell kompatibel,d.h. liegt sie wie die Helices in Abb.26 zum Teil in einem"verbotenen" Bereich und zum Teil in einem sterisch weiterhin zugelassenen, so wird sie entsprechend verkürzt. Es werden dann neue Werte für A_1,A_2 und die Länge errechnet und eingesetzt; die Stabilität der Helix, die an anderer Stelle im Programm gespeichert ist, wird ebenfalls neu aufaddiert. Gleichzeitig mit diesen Veränderungen werden die noch kompatibeln und zum Teil verkürzten Helices im Speicher-Vektor zum Anfang hin verschoben, wie dies in Abb.27 wiedergeben ist. Muss dann nach Auswahl der nächsten, in die Faltung aufzunehmenden Helix,erneut die Liste der verbleibenden Helices durchgegangen werden, um diejenigen Helices zu finden, die mit dem 2ten Kandidaten kompatibel sind, so braucht

das Programm den Vektor nur bis zum letzten der zusammengerückten Elemente durchzugehen. Dadurch

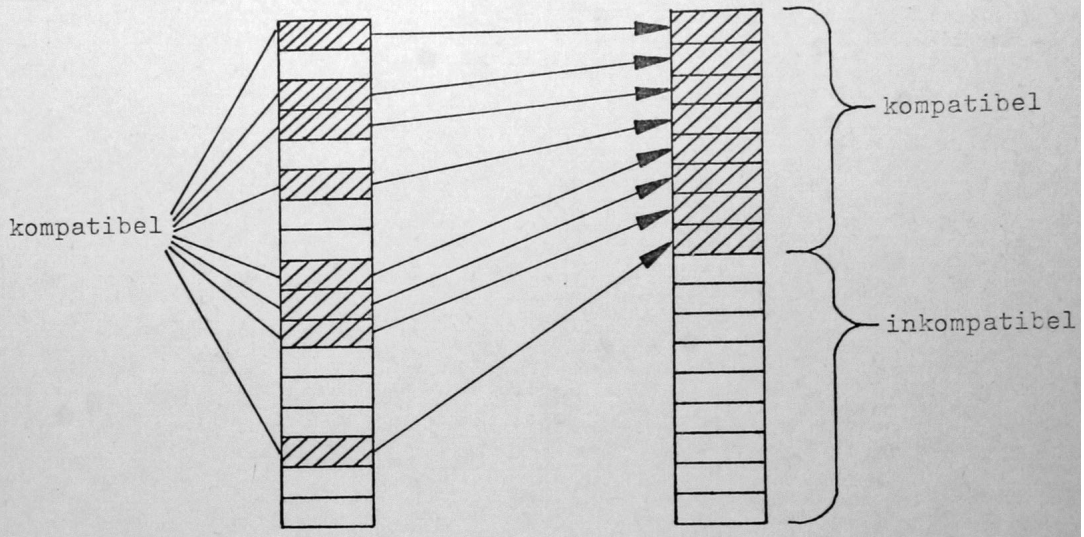


Abb.27:Zusammenrücken der zu einer Auswahlhelix kompatiblen Helices im Speichervektor des Programms SIM.FALTUNG

wird die Zeit, die das Programm für einen "Durchgang" durch diesen Vektor braucht gegen Ende des Programms immer kürzer. Im Programm FALTUNG war die Zeit für einen Durchgang durch die beiden Bindungsmatrices B und C am Ende so lang wie am Anfang des Algorithmus, dieser Nachteil ist in SIM.FALTUNG behoben. Im Abb. 28 ist der gesamte Ablauf des Programms SIM. FALTUNG in Form eines Flussdiagrammes dargestellt.

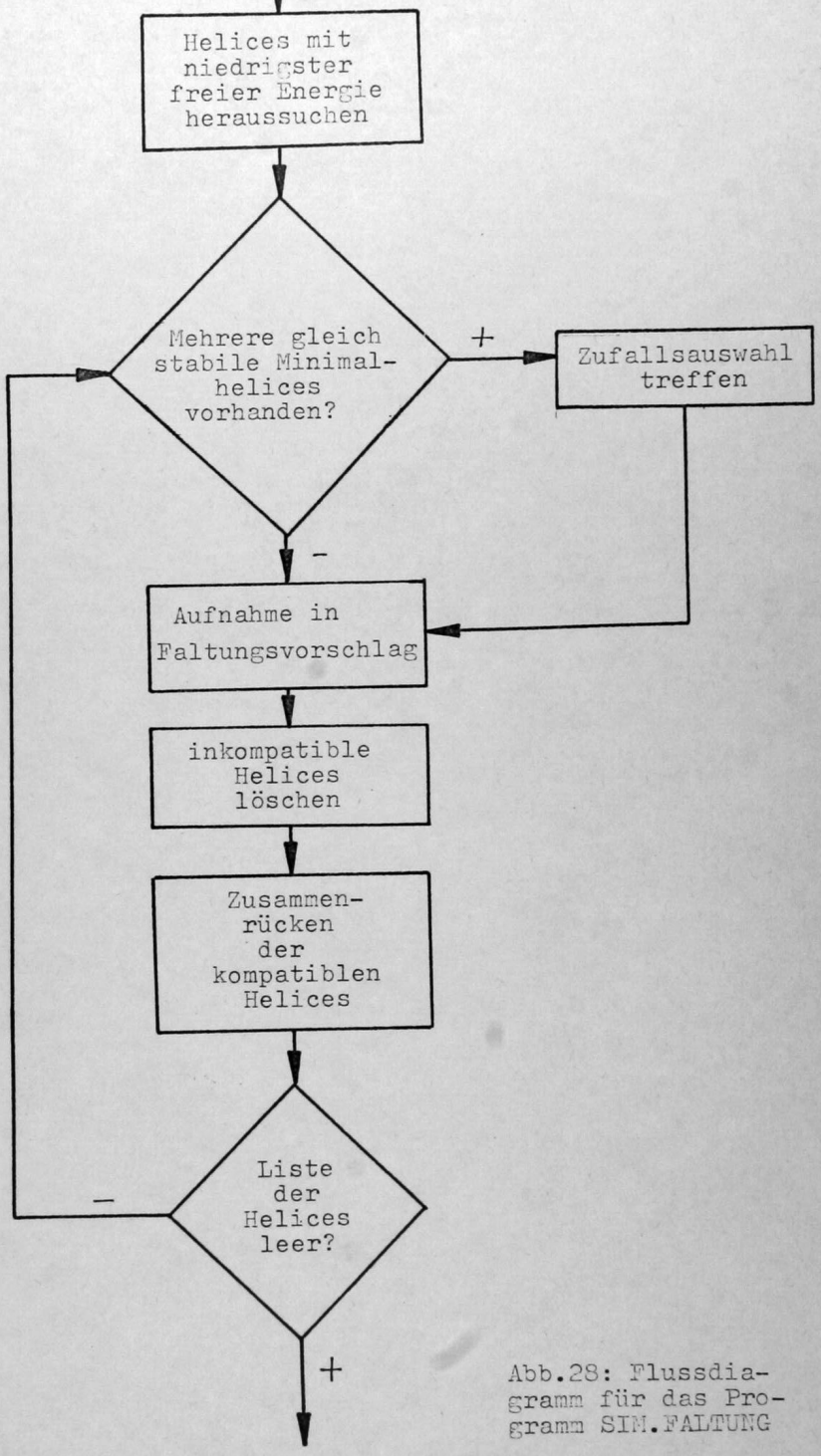


Abb.28: Flussdiagramm für das Programm SIM.FALTUNG

4.3.3 Prinzipien der Berechnung der freien Energie eines RNS-Moleküles mit hypothetischer Sekundärstruktur

Die Berechnung der freien Energie einer gefalteten RNS erfolgt im wesentlichen nach TINOCO, BORER et al. (1973) (37) . Die freien Energien der einzelnen Helices gehen in die Gesamt-freie Energie des Moleküles über. Zu dieser negativen stabilisierenden freien Energie werden die positiven destabilisierenden freien Energien der zwischen den gebundenen RNS-Teilen liegenden ungebundenen Abschnitte hinzugezählt. Diese ungebundenen RNS-Teile werden nach obiger Quelle in drei prinzipiell verschiedene Arten unterteilt, in Haarnadelschleifen (hairpin loops), Ausbauchungs- oder Bauchungsschleifen (bulge loops) und innere Schleifen (internal loops). Die folgende Abbildung gibt eine Vorstellung von der Definition dieser Bezeichnungen:

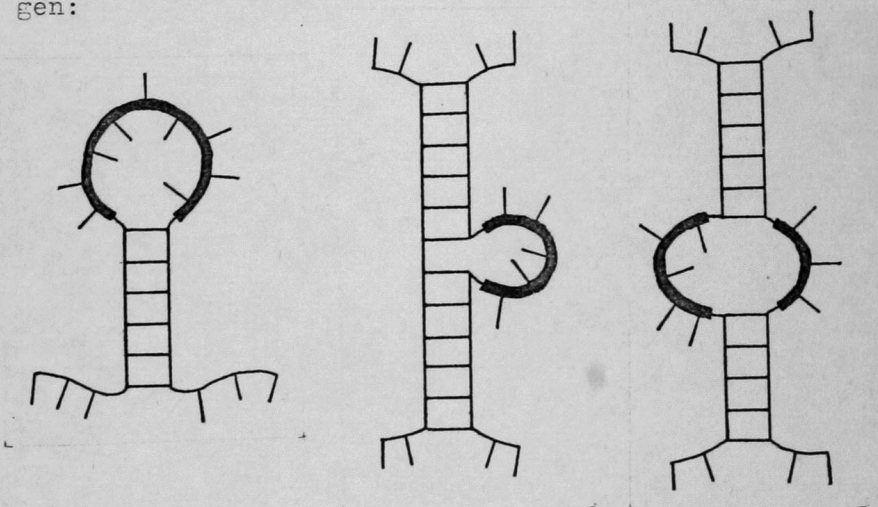


Abb. 29: Definition der Art der ungebundenen Molekülteile: Die Nucleotide in den verdickt gezeichneten Teilen entsprechen links einer Haarnadelschleife, in der Mitte einer Bauchungsschleife und rechts einer inneren Schleife.

Eine Bauchungsschleife bildet sich also, wenn zwei Helices auf einem der antiparallelen Stränge in Kontakt kommen,d.h., wenn E_1 der einen

Helix gleich A_1-1 der anderen Helix ist, oder wenn E_2 gleich A_2+1 wird. Eine innere Schleife bildet sich, wenn eine Helix sich in der Haarnadelschleife einer anderen Helix bildet, ohne mit dieser in Kontakt zu geraten. Entsprechend diesen drei Kategorien sind in der genannten Literatur verschiedene positive Werte für die freie Energie angegeben, die in der folgenden Tabelle wiedergegeben sind:

Anzahl der ungebundenen Basen	ΔG in kcal(±1kcal)		
	Innere Schleifen		
2 - 6	+ 2.0		
7 - 20	+ 3.0		
m(> 20)	+ 1.0 + 2.0·$\log_{10}$(m)		
	Bauchungsschleifen		
1	+ 3.0		
2 - 3	+ 4.0		
4 - 7	+ 5.0		
8 - 20	+ 6.0		
m(> 20)	+ 4.0 + 2.0·$\log_{10}$(m)		
	Haarnadelschleifen geschlossen durch		
	(G·C)	oder	(A·U)
3	+ 8.0		> 8.0
4 - 5	+ 5.0		+ 7.0
6 - 7	+ 4.0		+ 6.0
8 - 9	+ 5.0		+ 7.0
10 - 30	+ 6.0		+ 8.0
m(> 30)	3.5 + 2.0·$\log_{10}$(m)		5.5 + 2.0·$\log_{10}$(m)

Wenn man die freie Energie einer hypothetischen Sekundärstruktur nach diesen Angaben berechnen will, stösst man auf ungebundene Abschnitte, die sich in keine der obengenannten Kategorien einteilen lassen. Beispiele für solche Abschnitte sind in Abb.30 dargestellt. Abschnitte solcher Art werden weiterhin Zwischenstücke genannt und als Abkürzungen für die verschiedenen Arten ungebundener Sequenzen gelten:

HP = Haarnadelschleife (hairpin loop)
IL = innere Schleife (internal loop)
HL = Bauchungsschleife (bulge loog)
ZW = Zwischenstück

Die Berechnung der freien Energie der Zwischenstücke, die der Vollständigkeit halber erfolgen muss, wird entsprechend der Berechnung der inneren Schliefen vorgenommen.

Bei den Haarnadelschleifen wurden Schleifen, die (G·U) geschlossen wurden, willkürlich so behandelt, wie solche, die mit(A·U) geschlossen wurden. Auch muss eine Haarnadelschleife, die mit (G·U) geschlossen wird mindestens 4 Basen umfassen, wie dies bei (A·U)-Schleifen ebenfalls vereinbart wurde.

Die Gesamt-freie Energie der betreffenden Faltung ergibt sich aus der Summe der negativen stabilisierenden freien Energien der helicalen Bereiche und der positiven destabilisierenden freien Energien der ungebundenen Abschnitte der RNS. Je negativer als die Gesamt-freie Energie ist, desto stabiler wird die betreffende Molekülkonformation sein.

Die Stabilitätsberechnung nach obiger Tabelle wurde sowohl in das Programm FALTUNG inkorporiert, wie auch in das Programm SIM. FALTUNG.

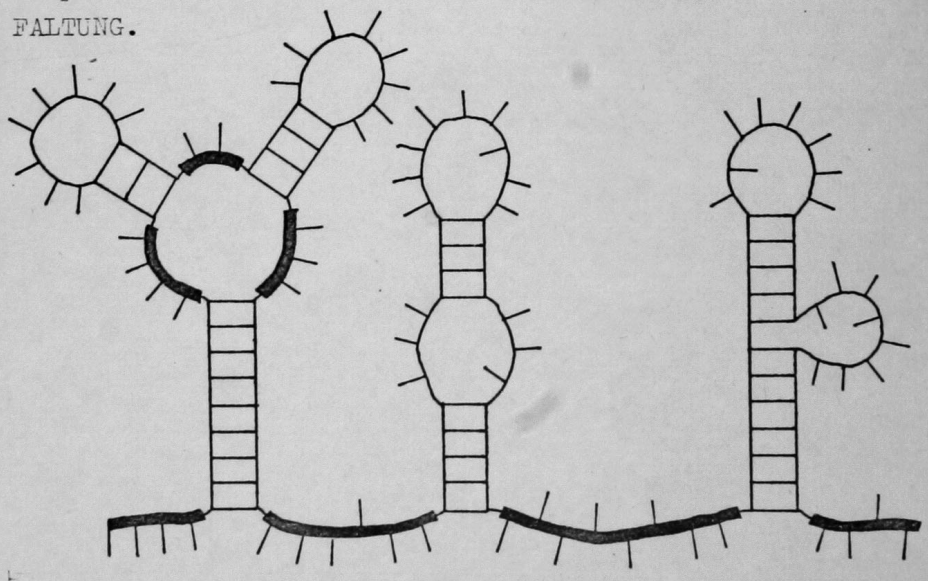


Abb.30: Dunkle Abschnitte = Zwischenstücke

4.3.4 Ergebnisse der einfachen Optimierung bei MS2, ϕX174 und einem Zufallsmessenger

Mit Hilfe der Programme FALTUNG und SIM.FALTUNG wurden für Abschnitte der Bakteriophagen MS2 und ϕX174 sowie eines Zufallsmessengers hypothetische Sekundärstrukturen ermittelt und ausgewertet.

4.3.4.1 Variation der Länge des optimierten Bereiches (FALTUNG)

Vornehmlich, um die obere Grenze des durch FALTUNG zu bearbeitenden Bereiches feststellen zu können, wurde eine Serie von immer länger werdenden Abschnitten aus MS2 getestet. Es handelte sich immer um das 5'-Ende des Bakteriophagen, das jeweils um 10 Nucleotide erweitert wurde, bis zu einer Maximallänge von 200 Nucleotiden. Die Abb.31 gibt die Bearbeitungsdauer und den Bedarf an Speicherplatz im Computer wieder, für die ersten 12 Programmläufe. Aufgrund der quadratischen Zunahme der Grösse der Bindungsmatrix nimmt auch die Grösse des für das Programm benötigten Speicherplatzes quadratisch zu. Die Bearbeitungsdauer ist direkt abhängig von der Grösse der vorhandenen Matrices und zeigt dasselbe Verhalten. Diese Abhängigkeiten machen eine Verwendung des Programms für Abschnitte mit einer Länge von mehr als 200 bis 300 Nucleotiden unrentabel. An eine Bearbeitung der Gesamtlänge des Bakteriophagen MS2 mit 3569 Nucleotiden ist nicht zu denken. Die obere Hälfte von Abb.32 zeigt die Veränderung des Prozentsatzes an ungebundenen Basen mit der steigenden Länge des vom Programm bearbeiteten Molekülteiles. Nach anfänglichen Schwankungen pendelt sich der Prozentsatz bei etwa 35% ein. Man kann also davon ausgehen, dass das 5'-Ende von MS2 , auch wenn es sich nur lokal falten kann, d.h. nicht mit anderen Bereichen des Moleküles in Kontakt treten kann, mindestens 65% helicale Bereiche aufweisen wird. Eine weitere interessante Beobachtung ergab sich bei Vergleich der

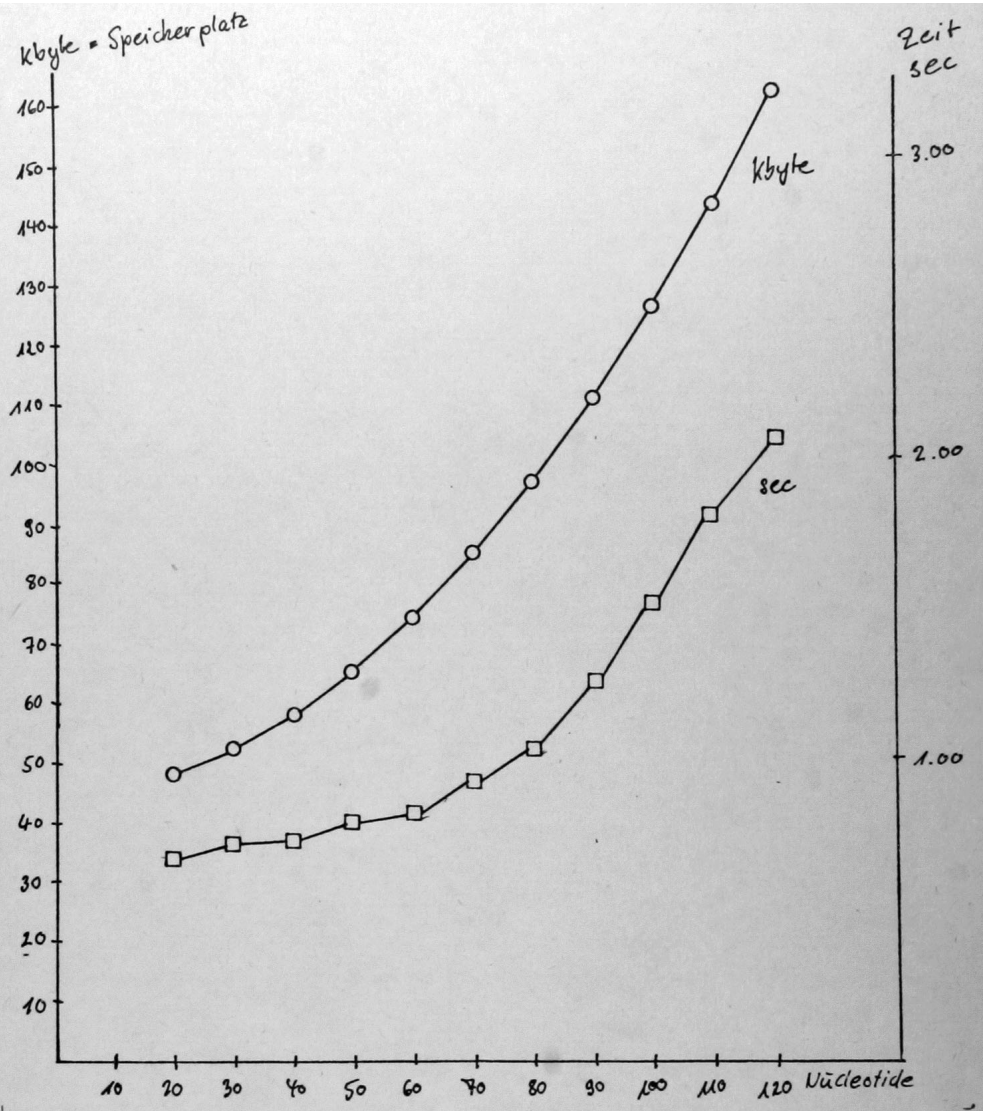


Abb.31: Veränderung der Bearbeitungsdauer und des Speicherplatzbedarfes im Programm FALTUNG mit zunehmender Länge des bearbeiteten Sequenzabschnittes. Kreise = Kbyte = Speicherplatz, Quadrat = Sekunden = Bearbeitungsdauer.

Basenzusammensetzung der Gesamtmoleküle mit der der ungebundenen Bereiche. Es zeigte sich ein erhöhter Gehalt an Uracil und ein erniedrigter Gehalt an Guanin in den freien Bereichen im Gegensatz zu den Gesamtmolekülen. Der untere Teil von Abb.32 zeigt die Variation der

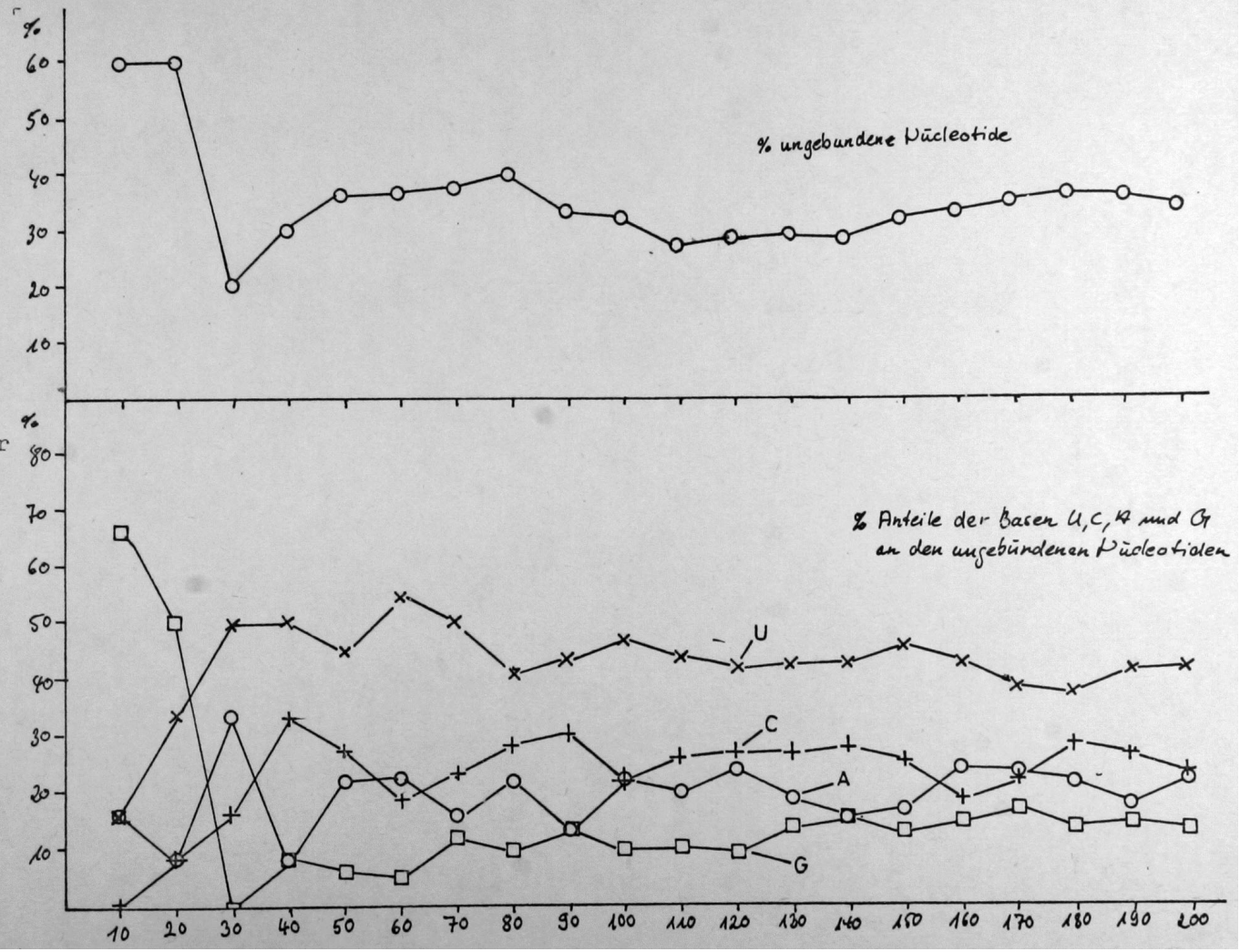


Abb.32: Oben: Veränderung des prozentualen Anteiles der ungebundenen Basen an der Gesamtzahl der Basen.
Unten: Veränderungen der prozentualen Anteile der einzelnen Basen an der Anzahl der ungebundenen Basen.

Prozentsätze der einzelnen Basen in den ungebundenen Bereichen mit der Länge des beobachteten Moleküles. Wieder ergaben sich anfängliche Schwankungen. Ab einer Bearbeitungslänge von 50 bis 60 Nucleotiden ist der U-Anteil deutlich an höchsten und der G-Anteil deutlich an niedrigsten. Diese Beobachtung liess sich auch durch statistische Auswertung der Faltung mit M = 200 bestätigen:
Diese Faltung enthielt 57 U,53 C,35 A und 55 G. Hiervon waren jeweils 28 U, 16 C 15 A und 9 G ungebunden. Hieraus errechnet sich:

p(U) = 0.285
p(C) = 0.265
p(A) = 0.175
p(G) = 0.275

Mit Hilfe dieser Wahrscheinlichkeiten kann man die Erwartungswerte der Anzahlen der einzelnen Basen in den ungebundenen Bereichen berechnen:

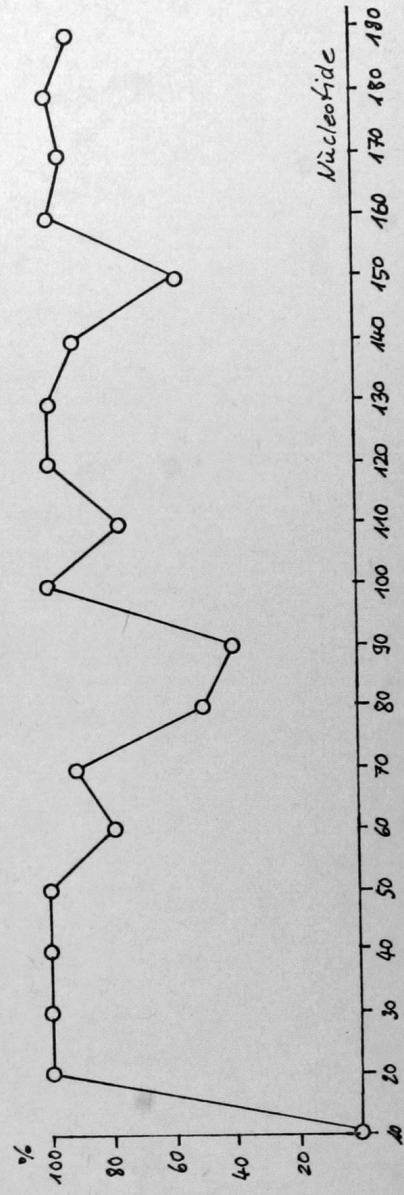


Abb.33: Ein Teil der Basenpaare einer Faltung stimmt mit den Basenpaaren in der nächsthöheren Faltung überein. Eingetragen sind die jeweiligen Prozentsätze dieser Basenpaare an den Gesamtanzahlen der Basenpaare.

$E_{ungeb.}(U) = 19.38$ $E_{ungeb.}(C) = 18.02$ $E_{ungeb.}(A) = 11.90$

$E_{ungeb.}(G) = 18.70$

Die statistische Analyse (u-Test) führt zu folgenden Werten:

	U	C	A	G
u =	2.316	-0.555	0.989	-2.634
Pr(u) =	0.9897	0.2894	0.8388	0.0042

Hiernach können also die Steigerung des U-Gehaltes und die Verrungerung des G-Gehaltes als statistisch signifikant angesehen werden. Die Variation der beiden anderen Basen ist nicht signifikant. Eine Weitere Beobachtung ergibt sich aus dem qualitativen Vergleich der gefalteten RNS-Abschnitte. Es gibt in einer Faltung zumeist einen Anteil an Basenpaaren, der sich in der um 10 Nucleotide verlängerten Faltung wiederfindet. In Abb. 33 sind die Anteile dieser Basenpaare an den jeweiligen Gesamtanzahlen von Basenpaaren eingetragen. Es zeigt sich, dass zwar oft die gesamten Helices einer kürzeren in die nächstlängere integriert werden, es aber auch zu einer drastischen Reduktion des Anteiles beibehaltener Helices kommen kann. Dies ist

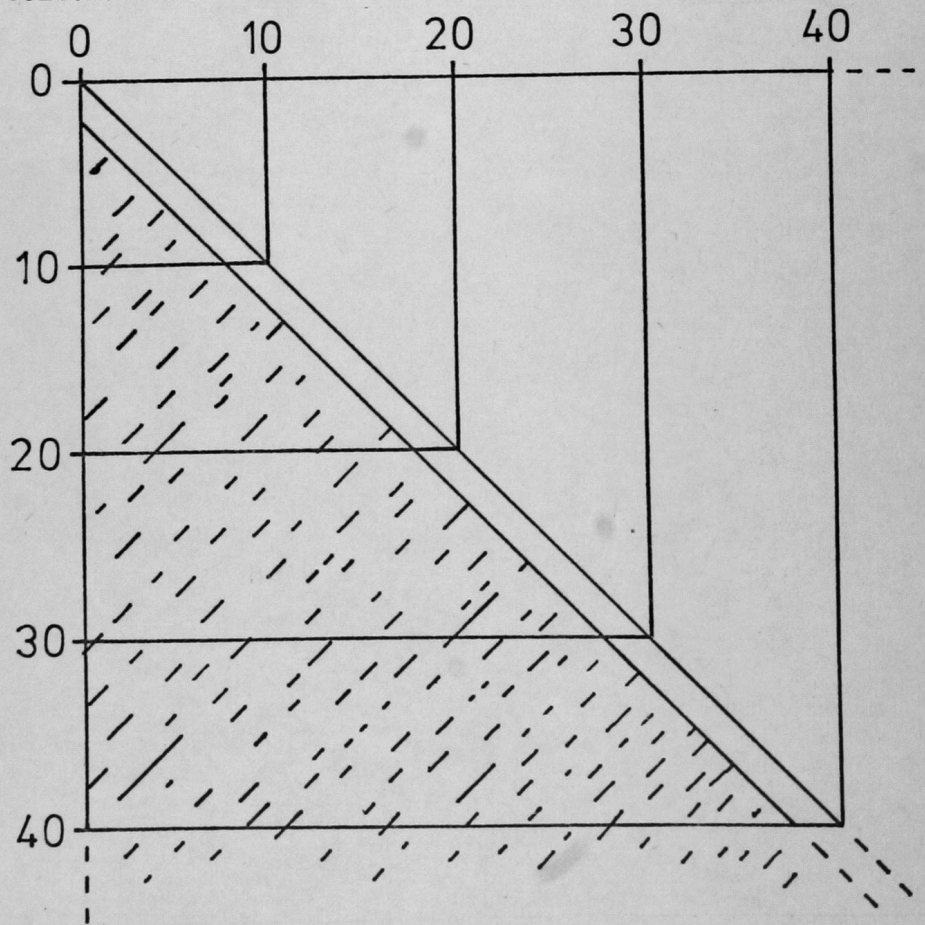


Abb.34: Bei Vergrösserung des bearbeiteten Sequenzabschnittes können sich auch stabile Helices zwischen den bereits gefalteten Abschnitten und den neu hinzugekommenen bilden. Dies verdeutlich sich an der oben gezeigten Bindungsmatrix.

immer dann der Fall, wenn durch das Grösserwerden der Bindungsmatrix auch stabile Helices in Bereichen auftauchen, in denen es bislang keine stabilen Helices gab, wie dies durch Abb. 34 verdeutlicht wird.
Für alle generierten Faltungen wurde auch die freie Energie berechnet. In Abb. 35 ist das Verhältnis der destabilisierenden positiven freien Energien zu den stabilisieren freien Energien wiedergegeben, in Abhängigkeit von der Länge der RNS-Abschnitte. Ab einer Länge von 70 Nucleotiden bleibt dieses Verhältnis bei etwa -2.0 konstant.
Abb. 36 gibt die Gesamt-freie Energie wieder, die für alle Faltungen durch die jeweilige Anzahl der betrachteten Nucleotide geteilt wurde. Dies ist also die Freie Energie pro Base. Wiederum ab etwa 70 Nucleotiden bleibt dieser Wert konstant bei ungefähr -1.2 kJ/Base. Wie zu erwarten verhalten sich besonders bei niedrigen M-Werten die beiden letztgenannten Funktionen reziprok. Gibt es mehr positive freie Energien als negative, so sinkt die Stabilität pro Base und umgekehrt.

Abb.35: Verhältnis der destabilisierenden freien Energien der ungebundenen Molekülteile zu der stabilisierenden freien Energie der Helixabschnitte in Abhängigkeit von der Faltungslänge

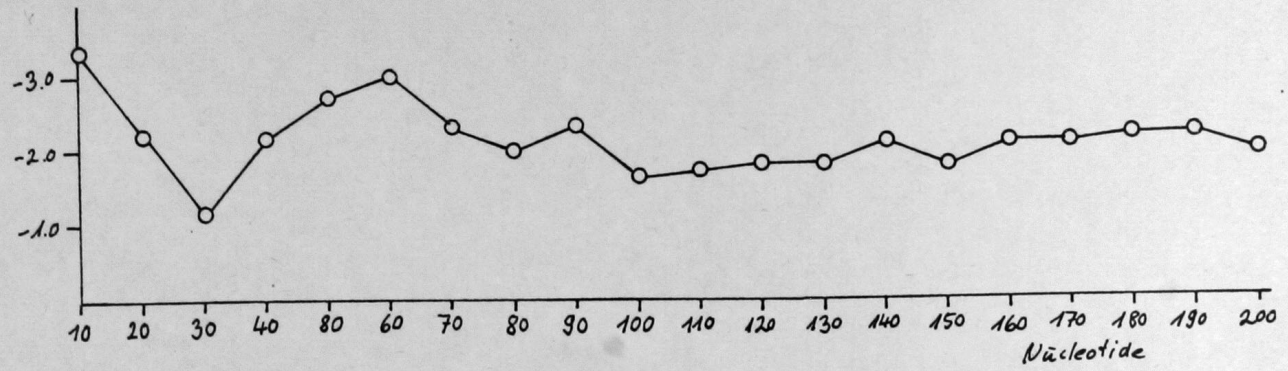
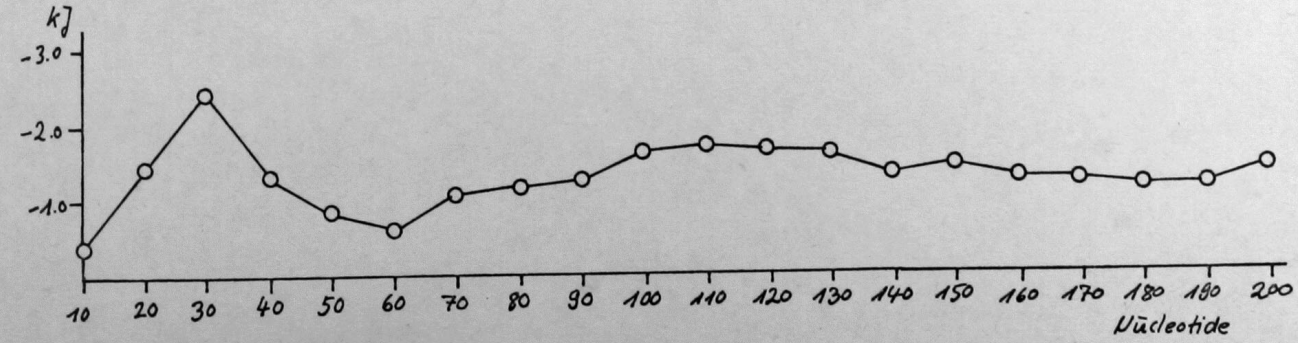


Abb.36: Gesamt-freie Energie pro Base in Abhängigkeit von der Faltungslänge



4.3.4.2 Vergleich der Faltungen der Längen 50, 100 und 200 von MS2

Es fragt sich nun, inwieweit die Unterschiede, die in Abschnitt 4.3.4.1 zwischen kürzeren und längeren gefalteten Abschnitten von MS2 aufgetreten sind, nur zufälliger Natur sind, also speziell nur bei dem betrachteten Anfangsstück von MS2 vorkommen oder allgemein für das Gesamtmolekül gültig sind. Auch ist es möglich, dass das Gesamtmolekül Unterschiede zwischen den Faltungen verschiedener Längen aufweist, die sich am Anfangsstück nicht nachweisen lassen. Um diesem Problem nachzugehen, wurden drei Testserien durchgeführt, die eine grössere Anzahl an gefalteten RNS-Abschnitten lieferten und somit einen statistischen Vergleich ermöglichen. Im ersten Schritt wurden alle nebeneinanderliegenden Abschnitte mit der Länge von 50 Nucleotiden von MS2 gefaltet. Dies ergibt bei der Länge von 3569 Nucleotiden dieses Messengers 71 verschiedene statistisch voneinander unabhängige Faltungen. Auf dieselbe Weise wurden sodann alle 35 Abschnitte zu je 100 Nucleotiden als zweite Testserie gefaltet und alle 17 Abschnitte zu je 200 Nucleotiden als dritte Testserie. Die folgende Tabelle gibt die Prozentsätze der ungebunden gebliebenen Nucleotide in den verschiedenen Testserien mit den dazugehörigen prozentualen Standardabweichungen an:

Länge	% ungebunden	% Standardabweichung
50	42.91	9.88
100	39.77	5.35
200	37.47	3.24

Mit diesen Daten lassen sich die Prüfwerte für einen statistischen Test (t-Test) errechnen, der uns die Möglichkeit gibt, zu entscheiden, ob die Mittelwerte zweier Stichproben sich signifikant unterscheiden. Als Signifikanzgrenze gilt wie bisher (siehe 4.1.5) 5%.

Vergleich:	t-Wert	Signifikanz?	Freiheitsgrad
50-100	1.755	+	104
100-200	1.628	-	50
50-200	2.233	+	86

Wann ein t-Wert eine Signifikanzgrenze überschritten hat, lässt sich leicht anhand von statistischen Tabellen überprüfen. Die Abnahme des Anteiles der ungebundenen Nucleotide in der Gesamtfaltung mit zunehmender Länge ist bei dem Vergleich 50 zu 100 und 50 zu 200 signifikant. Je länger also der vom Programm bearbeitete Abschnitt ist, desto mehr erhöht sich der Anteil gepaarter Nucleotide. Dieser Anteil lässt sich aber mit einer Verlängerung über 100 hinaus nicht mehr signifikant steigern.

Diese Unterschiede im Ausmass der Basenpaarbindungen können u.U. auf qualitative Unterschiede der Anzahlen ungebundener U,C,A oder G zurückgeführt werden. Zeigen sich also Unterschiede in den Prozentsätzen der einzelnen Basen in den ungebundenen Abschnitten bei Faltungen verschiedener Längen? Zunächst seien diese Prozentsätze für die einzelnen Faltungslängen sowie deren prozentuale Standardabweichungen angegeben:

Länge	Base	% ungebunden	% Standardabweichung
50	U	11.97	5.70
50	C	10.54	5.74
50	A	13.75	5.83
50	G	7.49	4.78
100	U	10.94	3.93
100	C	9.49	3.86
100	A	12.51	4.00
100	G	6.83	3.65
200	U	10.65	2.98
200	C	9.00	2.32
200	A	12.03	3.13
200	G	5.76	1.42

Diese Prozentsätze sind Anteile an der jeweiligen Gesamtzahl der Nucleotide. Um festzustellen, ob signifikante Unterschiede zwischen Anzahlen ungebundener U,C,A oder G in den verschiedenen Faltungen auftreten wurde wiederum ein t-Test durchgeführt:

Vergleich:	Base:	t-Wert:	Signifikanz?	Freiheitsgrad:
50-100	U	1.984	-	104
50-100	C	0.975	-	104
50-100	A	1.126	-	104
50-100	G	0.724	-	104
100-200	U	0.274	-	50
100-200	C	0.477	-	50
100-200	A	0.439	-	50
100-200	G	1.154	-	50
50-200	U	0.926	-	86
50-200	C	1.079	-	86
50-200	A	1.172	-	86
50-200	G	1.470	-	86

Eine Signifikanz trat in keinem Fall auf, man kann also davon ausgehen, dass keine Base nsorte bei Verlängerung des gefalteten Abschnittes ihren Anteil an den ungebundenen Basen signifikant ändert. Die Basenzusammensetzung der ungebundenen Abschnitte bleibt also unverändert.

Die folgende Tabelle führt die errechneten Mittelwerte der freien Energien der Faltungen verschiedener Länge auf, getrennt nach negativer freier Energie der helicalen Abschnitte und positiver freier Energie der ungebundenen Abschnitte sowie deren Summen:

Länge	ΔG	Mittelwert kJ	Standardabweichung kJ
50	negativ	-113.644	30.720
50	positiv	88.308	17.172
50	gesamt	-26.722	31.939
100	negativ	-250.568	38.124
100	positiv	153.907	24.471
100	gesamt	-96.709	43.532
200	negativ	-521.862	53.659
200	positiv	309.727	39.982
200	gesamt	-212.131	73.356

Diese Werte sind direkt noch nicht statistisch vergleichbar, deshalb wurden sie umgerechnet in freie Energie pro Base:

Länge	ΔG	Mittelwert kJ	Standardabweichung kJ
50	negativ	-2.273	0.614
50	positiv	1.766	0.343
50	gesamt	-0.534	0.639
100	negativ	-2.506	0.391
100	positiv	1.539	0.245
100	gesamt	-0.967	0.435
200	negativ	-2.609	0.268
200	positiv	1.549	0.200
200	gesamt	-1.061	0.367

Hierbei ist: $\Delta G(\text{gesamt}) = \Delta G(\text{negativ}) + \Delta G(\text{positiv})$

Um zu untersuchen, ob die Stabilität eines Moleküles bei Verlängerung des bearbeiteten Abschnittes signifikant zunimmt wurde ein t-Test durchgeführt, der sowohl die negative, wie auch die positive und die Gesamt-freie Energie der einzelnen Testserien miteinander vergleicht:

Vergleich:	ΔG	t-Wert:	Signifikanz?	Freiheitsgrad:
50-100	negativ	2.044	+	104
50-100	positiv	6.972	+	104
50-100	gesamt	6.223	+	104
100-200	negativ	1.231	+	50
100-200	positiv	0.140	-	50
100-200	gesamt	0.762	-	50
50-200	negativ	3.112	+	86
50-200	positiv	3.543	+	86
50-200	gesamt	4.612	+	86

Bei Steigerung der Test-Ketten-Länge von 50 auf 100 Nucleotide nehmen die Werte der negativen, positiven und Gesamt-freien Energie ab, dies zeigt sich aber sich bei einer Steigerung von 100 auf 200 Nucleotide nur noch bei der negativen freien Energie. Zusammenfassend kann festgestellt werden, dass bei einer Steigerung der Faltungslänge von 50 auf 100 der Anteil der gebundenen

Basen zunimmt und auch die thermodynamische Stabilität des Moleküls sich verbesert. Bei einer Steigerung von 100 auf 200 hingegen zeigt sich nur noch eine leichte Verbesserung der thermodynamischen Stabilität. Eine von der Länge abhängige Veränderung der Zusammensetzung der Basen in den ungebundenen Bereichen liess sich nicht feststellen.

4.3.4.3 Vergleich von Faltungen der Länge 200 bei MS2 und Faltungen der gleichen Länge bei einem Zufallsmessenger

Es gibt Computerprogramme, die Serien von zufällig verteilten Zahlen produzieren. Ein solches Programm wurde von mir benutzt, um ein RNS-Molekül zu generieren, dessen Basen gleichverteilt sind, also pU = pC = pA = pG = 0.25 ist. Es wurde, wie dies auch in den bisherigen, von mir erstellten, Computerprogrammen geschehen ist, jeder Base eine Zahl zugeordnet, nämlich U ≙ 1, C ≙ 3, A ≙ 5 und G ≙ 7 . Das Zufallsprogramm stellte nun eine Folge von diesen Zahlen her, bei der jede Zahl eine gleiche Wahrscheinlichkeit besitzt, an irgendeiner Stelle aufzutreten. Von diesem Zufallsmessenger, der im Anhang aufgelistet ist, wurden nun 21 Faltungen zu je 200 Nucleotiden erstellt und ausgewertet. Es wurde gezählt wieviel Basen gebunden und wieviel ungebunden vorlagen:

Länge	% ungebunden:	% Standardabweichung
200	40.71	3.10

Mit Hilfe des t-Tests wurde dieser Wert mit dem entsprechenden Wert der 200er Faltungen von MS2 verglichen:

Vergleich:	t-Wert:	Signifikanz?	Freiheitsgrad:
200_{MS2}-$200_{ZufallsRNS}$	1.948	+	36

MS2 zeigt also mit 37.47% ungebundenen Nucleotiden eine signifikante Erniedrigung gegenüber den Faltungen des Zufallsmessengers. Es ist nicht sehr sinnvoll, die Basenzusammensetzung der ungebundenen Abschnitte der Faltungen der ZufallsRNS mit den entsprechenden Werten von MS2 zu vergleichen, da MS2 eine leicht unterschiedliche Gesamtbasenzusammensetzung hat. Dennoch sollen diese Werte für den Zufallsmessenger hier

aufgelistet werden, um einerseits als Test für die Zuverlässigkeit des verwendeten Zufallsprogrammes zu dienen, also um zu zeigen, dass sich wirklich eine Gleichverteilung mit einer zu erwartenden Genauigkeit eingestellt hat und andererseits,dass auch bei den Faltungen des Zufallspolymers dieselben Verschiebungen der Basenzusammensetzung in den ungebundenen Abschnitten erfolgt, wie auch bei den Faltungen von MS2:

Länge:	Base:	% der Gesamtanzahl:	%Standardabweichung:
200	U	25.26	2.99
200	C	23.95	3.53
200	A	24.74	3.74
200	G	26.05	2.85

Länge:	Base	% ungebunden:	%Standardabweichung:
200	U	10.64	2.56
200	C	8.93	2.71
200	A	13.50	2.48
200	G	7.64	2.03

Wiederum treten U und A in den ungebundenen Bereichen am häugisten auf und C und G weniger häufig. Die freien Energien der Faltungen des Zufallspolymers weisen folgende Werte auf:

Länge:	ΔG	Mittelwert kJ	Standardabweichung kJ
200	negativ	-472.005	47.871
200	positiv	317.301	37.287
200	gesamt	-154.684	59.494

oder umgerechnet pro Base:

Länge:	ΔG	Mittelwert kJ	Standardabweichung kJ
200	negativ	-2.360	0.239
200	positiv	1.587	0.186
200	gesamt	-0.773	0.297

Die Stabilitäten der 200er von MS2 und vom Zufallspolymer wurden statistisch verglichen:

Vergleich:	ΔG	t-Wert:	Signifikanz?	Freiheits-grad:
200-200	negativ	3.025	+	36
200-200	positiv	0.061	-	36
200-200	gesamt	2.675	+	36

Somit ist also bei MS2, unter Anwendung desselben Faltungsalgorithmus, der Anteil gebundener Basen erhöht gegenüber RNS mit zufälligen Sequenzen, was mit einer signifikanten Erniedrigung der negativen und Gesamt-freien Enerige verbunden ist, jedoch keinen Unterschied in der destabilisierenden positiven freien Enerigie bedingt.

4.3.4.4 Einfache Optimierung von MS2

Mit einigen weiter unten beschriebenen Abwandlungen konnte das Programm SIM.FALTUNG zur Erstellung einer einfachen Optimierung des Gesamtmoleküles von MS2 eingesetzt werden. Bislang hat sich die Erstellung von Sekundärstrukturvorschlägen von RNS-Molekülen mit Hilfe des Computers auf kurzkettige Moleküle beschränkt. Hiermit wird erstmals die Möglichkeit aufgezeigt, auch bei längerkettigen RNS-Molekülen - MS2 hat 3569 Nucleotide - Gesamtfaltungen aufzustellen.

4.3.4.4.1 Beschreibung des speziellen Verfahrens zur Erstellung der Gesamtfaltung von MS2

Nach der Zählung aller möglichen Helices durch das Programm HELIX.LISTE sind in MS2 genau 367820 verschiedene Helices bei der Erstellung von Sekundärstrukturen in Betracht zu ziehen. Diese umfassen 1079069 verschiedene Basenpaarkombinationen. Will man nun das Programm SIM. FALTUNG mit dieser Menge von Helices arbeiten lassen, so stösst man computertechnisch auf Schwierigkeiten bezüglich des Speicherplatz-und Rechenzeitbedarfes. Darum wurden in einem ersten Schritt von dem Programm HELIX. LISTE die 9957 stabilsten Helices herausgesucht (das sind die Helices mit einer freien Energie kleiner als -10.0 kcal (bzw. -41.855 kJ)). Aus dieser Menge suchte dann das Programm SIM.FALTUNG 92 Helices heraus, nach dem Prinzip der Auswahl der stabilsten Elemente, wie oben (4.3.2.2) beschrieben. Es existiert dann aus den 9957 Ausgangshelices keine einzige mehr, die zu der Faltung der 92 ausgewählten Helices kompatibel wäre.
Im Weiteren wurde versucht, diese "Grobfaltung" durch Helices mit einer freien Energie grösser als -41.855 kJ zu ergänzen. Mit Hilfe von

HELIX.LISTE wurden dann die restlichen Helices, die in den ungebunden gebliebenen Bereichen noch möglich waren, ermittelt. HELIX.LISTE lässt sich nur dann nicht verwenden, wenn es sich um die Auffindung von Helices handelt, die zwischen zwei getrennten, ungebundenen Bereichen auftreten können, z.B. zwischen den beiden Hälften einer inneren Schleife oder zwei Zwischenstücken. Hierfür wurde eine neue Version von HELIX.LISTE geschaffen, die auch derartige Helices auffinden kann. Es handelt sich bei den genannten Kombinationen um rechteckige Ausschnitte aus der Gesamtmatrix von MS2. Es genügt, wenn man den Abtastvorgang, bzw. die schrittweise Veränderung der Indices i und j so verändert, dass er dem Verlauf des Pfeiles in Abb. 37 entspricht. Es muss bei

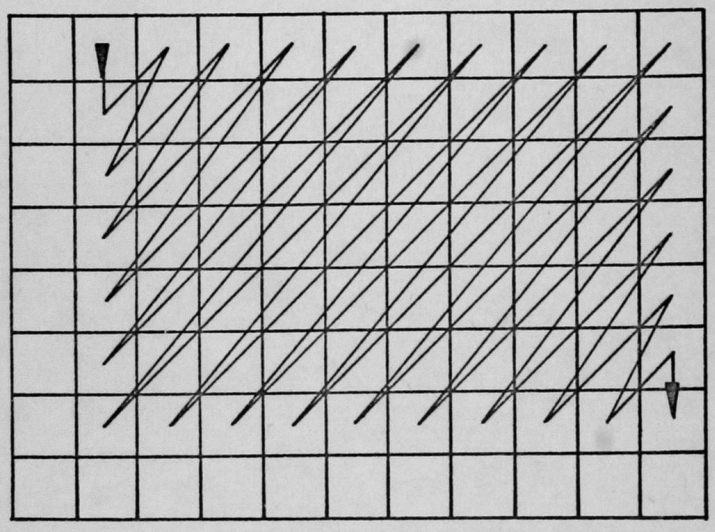


Abb.37:Rechteckiger Matrixausschnitt aus der Gesamtbindungsmatrix einer längerkettigen RNS. Um alle Helices in diesem Teil zu finden, müssen die Indices i und j so variiert werden, wie dies der Pfeil in der Abbildung wiedergibt.

diesem Verfahren der gesamte rechteckige Matrixausschnitt abgesucht werden. Man vergleiche dies mit dem ursprünglichen Abtastalgorithmus von HELIX.LISTE, der in Abb. 22 dargestellt ist.

Auf diese Weise konnten nun weitere 3413

Helices aufgefunden werden, die zu den bereits aufgefundenen 92 Helices kompatibel sind. Diese 3413 Helices wurden zusammen mit den 92 Helices wiederum als Eingabe für das Programm SIM.FALTUNG verwendet, das hiermit die Gesamtfaltung von MS2 erstellen konnte. Diese Faltung ist im hinteren Umschlagblatt beigefügt. Sie besteht aus 273 verschiedenen Helices, also aus 0.0742 % aller möglichen Helices. Diese 273 Helices stellen wiederum 0.1077 % aller möglichen Basenpaare dar. Und mit möglichen Basenpaaren sind nur solche gemeint, die in Helices mit mindestens einer Länge von 2 auftreten und sich in Faltungen einbauen lassen.

4.3.4.4.2 Vergleich der per Computer generierten Faltung mit dem Sekundärstrukturvorschlag der Arbeitsgruppe FIERS

Von der Arbeitsgruppe FIERS, der die Sequenzaufklärung von MS2 (8,9,10,11,12,30,31) zu verdanken ist, wurde ein Vorschlag für die Sekundärstruktur dieses Moleküls gemacht. Es ist nun möglich, diesen Vorschlag mit der in dieser Arbeit per Computer generierten Sekundärstruktur zu vergleichen. In der folgenden Liste sind einige Eigenschaften der beiden Strukturen einander gegenübergestellt:

	Per Computer generierte, hypothetische Sekundärstruktur	Sekundärstrukturvorschlag der Arbeitsgruppe FIERS
% gebundene Nucleotide	65.17	69.94
% freie Nucleotide	34.83	30.06
% U in ungeb. Teilen	27.29	32.43
% C " " "	24.64	18.83
% A " " "	32.21	35.88
% G " " "	15.86	12.86
% identische Basenpaare	27.13	25.28
Δ G negativ	-9931.26 kJ	-11527.70 kJ
Δ G positiv	4331.99 kJ	4938.89 kJ
Δ G gesamt	-5599.26 kJ	-6588.81 kJ

Beide Sekundärstrukturvorschläge sind im hinteren Umschlagblatt beigelegt.
Die Computerfaltung besitzt also etwa 4.77% weniger Basenpaarbindungen als die Faltung nach FIERS, ihre negative freie Energie ist numerisch grösser, was auch einen Anstieg der Gesamt-freien Energie bedingt, dies heisst, dass die Stabilität aller helicalen Bereiche für die Computerfaltung geringer ist, allerdings ist ihre Destabilisierung vorteilhafter. In der Faltung nach FIERS ist die negative freie Energie erniedrigt auf Kosten der destabilisierenden positiven freien Energie.

Die Basenverteilung in den beiden Faltungen zeigt dieselben Tendenzen. Das Vorherrschen von U und A in den ungebundenen Teilen ist allerdings bei der Faltung nach FIERS noch ausgeprägter. Etwa ein Viertel aller Basenpaare stimmen überein. Hierbei handelt es sich aber zumeist um Helices, die zwischen eng beieinanderliegenden Teilen des Moleküles gebildet werden, sie haben also "lokalen" Charakter, in so gut wie allen nicht-lokalen Fällen differieren die beiden Faltungen, was auch eine völlig unterschiedliche Gesamtstruktur bedingt. In der Computerfaltung kommen langgestrecktere Verästelungen vor als in der FIERSschen Faltung. Am auffallendsten ist das enge Zusammentreten der beiden Enden des Moleküles in der Computerfaltung, die beide durch eine gleich lange lokale Helix umgebogen werden, also etwa gleiche Form haben. Dies hat möglicherweise eine Bedeutung für die Funktion des Moleküls.
Zum weiteren Vergleich der beiden Faltungen ist in der folgenden Tabelle angegeben, wie oft ungebundene Teilsequenzen bestimmter Längen jeweils vorkommen:

Länge der ungeb. Teilsequenzen	Anzahlen der ungeb. Teilsequenzen in Computerfaltung:	Faltung nach FIERS:
1	123	199
2	70	83
3	53	53
4	62	44
5	41	18
6	12	18
7	19	14
8	10	4
9	1	0
10	5	2
11	2	2
12	0	1
13	0	1
15	0	1

Bei der Computerfaltung zeigen sich weniger kurze Abschnitte (1-2), dafür sind die längeren Abschnitte vermehrt (4-11), allerdings treten keine Abschnitte auf, die länger als 11 freie Nucleotide wären.

Aus den obigen Ergebnissen ist zu folgern, dass Faltungen, die, wie die Faltung nach FIERS, stabiler sind als die per Computer aufgestellte, wahrscheinlich einen noch geringeren Gehalt an G und C in den ungebundenen Bereichen besitzen und kürzere Zwischenstücke und Schleifen (loops) aufweisen. Es wäre nun möglich, den Faltungsalgorithmus dahingehend zu verbessern, dass er bevorzugt Faltungen mit den oben genannten Eigenschaften generieren würde. Dies könnte zwar zu Faltungen führen, die bessere thermodynamische Eigenschaften besässen als die hier vorgestellte, aber es wäre immer noch keine Garantie gegeben, dass man hiermit die bestmögliche Faltung für die betreffende RNS vorgeschlagen hätte.
Darum wird im folgenden Kapitel (4.4) ein Vorschlag für ein völlig anderes Verfahren zur Optimierung von Sekundärstrukturen gemacht, das ich "Völlige Optimierung" genannt habe und das die Schwierigkeiten und Unzulänglichkeiten der einfachen Optimierung überwinden soll.

4.4 Völlige Optimierung der Sekundärstruktur von RNS-Molekülen

Unter Weiterführung der bisherigen theoretischen Überlegungen wird versucht, einen Algorithmus zu finden, der in der Lage ist, die absolut stabilste Sekundärstruktur eines beliebigen RNS-Moleküles zu finden. Dieser Versuch ist darum so schwierig, weil die stabilste Sekundärstruktur eines RNS-Moleküles nicht unbedingt diejenige ist, die den höchsten Anteil an Basenpaarbindungen besitzt. Erhöht man die Anzahl der in einer bestimmten Faltung integrierten Helices, und verbessert dadurch den Wert der negativen freien Enerige, so erhöht man gleichzeitig auch die positive freie Energie des Moleküles durch die Vermehrung der ungebundenen Abschnitte. Es kann vorkommen, dass durch den Einbau einer Helix, die Gesamt-freie Energie einer Faltung steigt, also verschlechtert wird. Man muss hier versuchen einen Kompromiss zu finden, in dem Stabilisierung und Destabilisierung optimal zusammenwirken.

4.4.1 Warum ein neuer Ansatz?

Ausgehend von dem bisherigen Verfahren der einfachen Optimierung bzw. der Maximumwahl der Helices sind eine Reihe weiterer Optimierungsalgorithmen denkbar.Es liesse sich zum Beispiel das Auswahlprinzip so verändern, dass nicht mehr die längste Helix einer Reihe von Helices mit gleicher Stabilität ausgewählt würde, sondern diejenige mit der kleinsten freien Energie pro Base, sprich die Kürzeste der betrachteten Helices, hierdurch könnte u.U. die Gesamt-freie Energie pro Base erniedrigt werden. Oder es wäre denkbar, bei der Auswahl der Helices im Anfang solche Helices zu bevorzugen , die sich lokal bilden; diese Helices führen nur zu einer vergleichsweise geringen Reduktion der Menge der möglichen Helices als solche, deren beiden Teilstücke weiter auseinander liegen. Es ist bei letzten

Helices möglich, dass mehr als 50% aller möglichen Helices nicht mehr mit einer Faltung kompatibel sind. Dies ersieht man z.B. aus Abbildung 24 wo die Bereiche 5 und 6, in denen die nicht kompatiblen Helices liegen, bei einer nichtlokalen Helix mehr als die Hälfte der betrachteten Matrixhälfte ausmachen können. Auch wäre es möglich einem Algorithmus gezielt den Gehalt der ungebundenen Abschnitte an G und C zulassen verringern, in dem man bevorzugt Helices in die Faltung aufnimmt, die einen hohen G und C-Gehalt besitzen. Auch kann versucht werden, durch Aufnahme eng beieinander liegender Helices die Länge der ungebundenen Abschnitte zu verkürzen. Durch solche und andere Veränderungen am bisherigen Algorithmus würden vielleicht stabilere Faltungen als bisher aufgefunden werden, jedoch wäre keine Gewähr gegeben, ob es sich auch um die wirklich absolut stabilste mögliche Sekundärstruktur mit der nachweislich niedrigsten freien Energie handeln würde. Es gibt aber noch weitere Hinweise für die prinzipielle Unzulänglichkeit der bisherigen Algorithmen:

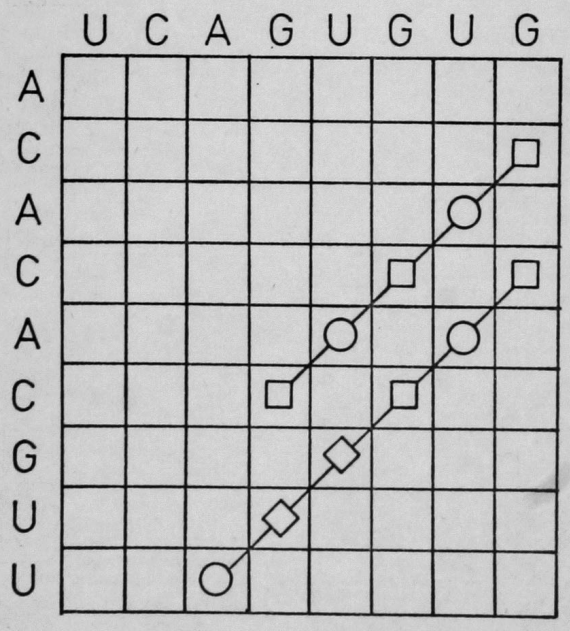
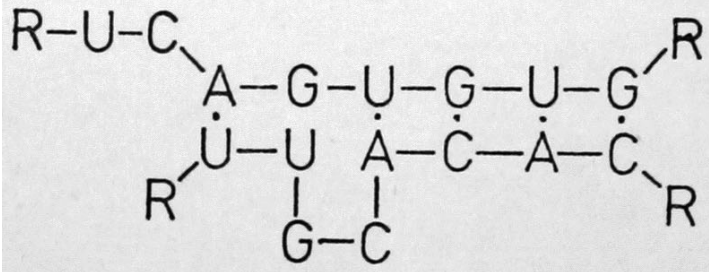
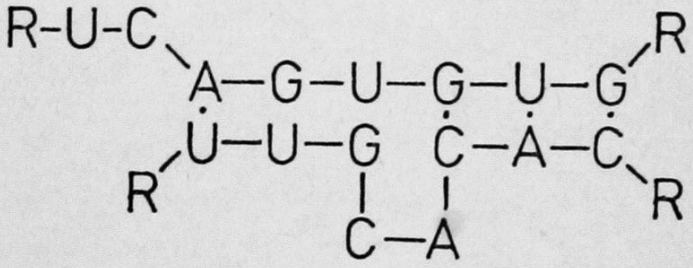
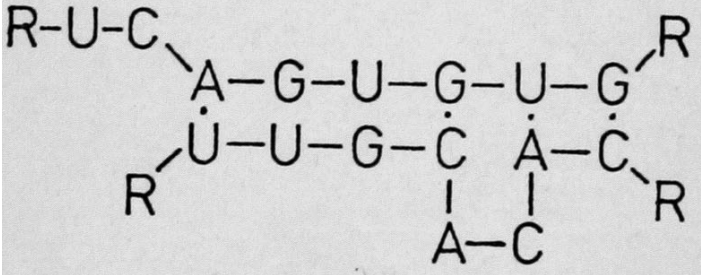


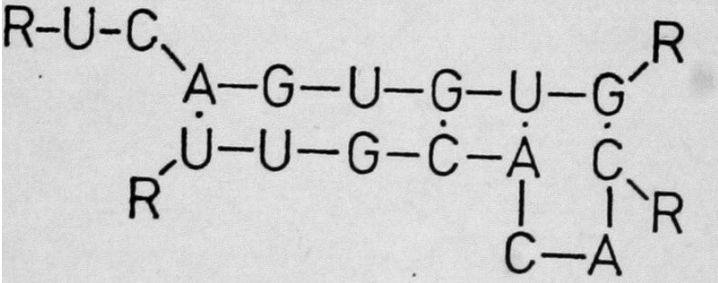
Abb.38:Beispiel für zwei einander stark überlappende Helices. Nicht alle möglichen Basenpaare in Bindungsmatrixausschnitt eingetragen.

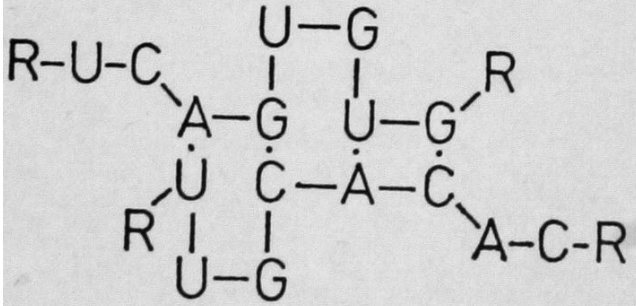
sich überlappenden Helices.











1. Bislang ist es nicht möglich, Teile von Helices in eine Faltung zu inkorporieren. Dies ist interessant in Fällen, in denen sich die Nucleotide zweier Helices teilweise miteinander überlappen. Z.B. treten bei der Kombination der Sequenz R-U-C-A-G-U-G-U-G-R mit der Sequenz R-C-A-C-A-C-G-U-U-R zwei einander fast ausschliessende Helices auf. Siehe Abbildung 38. Es gibt nun eine ganze Reihe von Faltungen, die Teile dieser beiden Helices enthalten, aber keine der beiden vollständig inkorporieren. In Abbildung 39 sind sechs von insgesamt zehn dieser Möglichkeiten wiedergegeben. (Mit R ist eine beliebige Fortsetzung der Nucleotidkette gemeint). Welche dieser Formen im Kontext des Gesamtmoleküles die energetisch günstigste ist, kann erst nach Berechnung der jeweiligen Gesamtstabilität gesagt werden. Es muss keineswegs der Fall sein, dass nur die Formen, in denen mindestens eine Helix vollständig enthalten ist, die günstigsten sind, aber nur solche konnten bislang mit den in FALTUNG und SIM. FALTUNG gegebenen Algorithmen gefunden werden. Sehr leicht kann durch eine Kombination von Teilen längerer Helices eine Faltung erreicht werden, die die bisher bekannten an Stabilität übertrifft.

2. Es kann weiterhin Kombinationen von weniger stabilen Helices geben, die gegenüber Kombinationen aus stabileren Helices energetisch günstiger sind. Eine Faltung aus einer sehr stabilen Helix und einer Reihe von kurzen, wenig stabilen Helices kann instabiler sein als eine Faltung aus einer Reihe von kompatiblen mässig stabilen Helices. U.U wird durch den Einbau einer Helix, z.B. mit einer freien Energie von 0.0 kJ die Gesamt-freie Energie der Faltung verschlechtert, es müssen also auch Faltungen auffindbar sein, in denen diese Helix zwar möglich aber nicht realisiert ist.
Es ist also nötig, einen Algorithmus zu schaffen, der Faltungen aus Teilhelices und mindersta-

bilen Helices auffinden und energetisch berechnen kann. Auch sollte er energetisch ungünstige Helices ausschliessen können. Nur ein solcher Algorithmus wird in der Lage sein, diejenige Faltung mit der niedrigsten freien Energie, das völlige Optimum, zu finden.
Ein Algorithmus, der diese Voraussetzungen erfüllt, wird in den folgenden Abschnitten abgeleitet und seine Realisierung durch das Programm MODELL beschrieben.

4.4.2 Ableitung des Suchalgorithmus und Berechnung der Höchstzahl möglicher Faltungen

In einem ersten Schritt muss ein Algorithmus gefunden werden, der in der Lage ist, sämtliche möglichen Faltungen einer RNS zu generieren. Der zweite Schritt besteht darin, Methoden zu finden, diesen Algorithmus, der wegen der grossen Zahl möglicher Faltungen sehr langwierig sein wird, auf geeignete Weise zu verkürzen, so dass ökonomisch mit dem Computer gearbeitet werden kann. In diesem Abschnitt wird nur auf den ersten Schritt eingegangen, der zweite ist dem Kapitel 4.4.3 vorbehalten.
Aus den Überlegungen des letzten Abschnittes ersieht man, dass Helices nicht als Ganzes die Basis für die Auffindung aller möglichen Faltungen bilden können, da auch Teile von Helices berücksichtigt werden müssen. Als Ausgangsmaterial können also nur die einzelnen Basenpaare dienen.
Es seien in einer RNS oder einem RNS-Abschnitt X genau L verschiedene Basenpaare (N·N) möglich. Als fortlaufende Indices dienen m und n mit: $1 <= m,n <= L$. Hiermit ist es möglich bei Durchnummerierung aller Basenpaare das m-te oder n-te Basenpaar mit $(N\cdot N)_m$ bzw. $(N\cdot N)_n$ zu bezeichnen. Es bedeute fernerhin

$$(N\cdot N)_m == (N\cdot N)_n$$

dass die Basenpaare m und n kompatibel sind, also

folgenden Überlegungen benötigt:

Mit $|\{\ldots\}|$ ist der Betrag der Menge $\{\ldots\}$ gemeint, d.h. die Anzahl seiner Elemente. Es ist z.B.

$$|\{(N\cdot N) \mid (N\cdot N) \text{ möglich in } X\}| = L$$

Eine Faltung ist eine Teilmenge der Menge aller möglichen Basenpaare, in der alle Basenpaare zueinander kompatibel sind. Wieviele solcher Teilmengen gibt es? Oder wieviele Faltungen sind bie einer gegebenen Ausgangsmenge von Basenpaaren möglich? Zunächst einmal gibt es L verschiedene Faltungen, in denen nur ein Basenpaar inkorporiert ist. Wieviele Faltungen mit 2 Basenpaaren kann es geben?

Es ist

$$|\{(N\cdot N) \mid (N\cdot N) == (N\cdot N)_m\}| <= L - 1$$

Das heisst, dass zu jeder der L verschiedenen Basenpaare höchsten L-1 verschiedene Basenpaare zur Erstellung einer Faltung mit 2 Basenpaaren verwendet werden können. Es sind genau L - 1 Basenpaare, wenn ein Basenpaar mit allen anderen kompatibel ist, und es sind 0 Basenpaare, wenn es mit keinem weiteren kompatibel ist. Somit ist die Anzahl aller Faltungen mit 2 Basenpaaren kleiner gleich $L\cdot(L-1)$. Entsprechend gibt es höchstens $L\cdot(L-1)\cdot(L-2)$ Faltungen mit 3 Basenpaaren.

Es gilt fernerhin:

$$L\cdot(L-1)\cdot(L-2) = \frac{L!}{(L-3)!}$$

Allgemein gibt es höchstens $\frac{L!}{(L-m)!}$ verschiedene Faltungen mit m Basenpaaren. Somit ist die Anzahl aller Faltungen mit höchstens m Basenpaaren nicht grösser als

$$\frac{L!}{(L-1)!} + \frac{L!}{(L-2)!} + \frac{L!}{(L-3)!} + \cdots + \frac{L!}{(L-m)!}$$

Bei der bislang als Beispiel abgehandelten 20 Nucleotide langen Anfangs-RNS von MS2 ist L = 64. Dies gibt die Anzahl von Basenpaaren an , die in der Bindungsmatrix dieses RNS-Abschnittes auftreten und Teile von Helices sind. Die längstmögliche Helix für dieses Beispiel hat 8 Basenpaare. Somit ist die Anzahl aller möglichen Faltungen mit höchstens 8 Basenpaaren kleiner als $1.8165\cdot10^{14}$. Wenn die wirkliche Anzahl der möglichen Faltungen nur gering unter dieser Anzahl liegt, dann ist ein systematisches Absuchen aller Faltungen sehr aufwendig.

Zum Auffinden aller möglichen Faltugen kann man zuerst alle Faltungen mit einem Basenpaar bilden, diese auswerten, dann sucht man zu jeder Faltung die Menge aller Basenpaare zusammen, die mit dem bereits inkorporierten Basenpaar kompatibel sind. Sodann wird jede 1er Faltung der Reihe nach mit jedem kompatiblen Basenpaar erweitert, so dass man alle möglichen 2er Faltungen erhält und auswerten kann. Aus der Menge der zu einem Basenpaar m kompatiblen Basenpaare wird also ein zweites Basenpaar n entnommen und wiederum festgestellt, welche Basenpaare von denjeingen, die mit R kompatibel sind auch mit S kompatibel sind. Diesen Vorgang wiederholt man repetitiv, bis keine weiteren Basenpaare mehr in die Faltungen inkorporierbar sind und hat dann alle möglichen Faltungen erhalten.

Dieses Suchverfahren lässt sich mit Hilfe eines baumförmigen Graphen verdeutlichen. Als Beispiel

sei ein Fall einer RNS genommen, in der vier Basenpaare $((N\cdot N)_1,(N\cdot N)_2,(N\cdot N)_3,(N\cdot N)_4)$ möglich sind, zwischen denen die folgenden Kompatibilitätsbeziehungen herrschen:

	$(N\cdot N)_1$	$(N\cdot N)_2$	$(N\cdot N)_3$	$(N\cdot N)_4$
$(N\cdot N)_1$	=/=	==	=/=	==
$(N\cdot N)_2$	==	=/=	==	==
$(N\cdot N)_3$	=/=	==	=/=	==
$(N\cdot N)_4$	==	==	==	=/=

Der zu diesen vier Basenpaaren gehörige Graph aller möglichen Faltungen ist in Abbildung 40 dargestellt. Von einem Startpunkt ausgehend verzweigt sich der Graph zunächst zu allen vier möglichen Basenpaaren (von denen nur die Nummern eingezeichnet sind). Jeder Kreis in der Abbildung wird Knotenpunkt genannt und stellt eine bestimmte Faltung dar. Die Anzahl der Knotenpunkte entspricht der Anzahl der möglichen Faltungen. Im ersten Verzweigungsschritt wurden also alle 1er Faltungen erreicht. Sodann wurde herausgesucht, welche Basenpaare jeweils zu den einzelnen Ausgangsbasenpaaren noch kompatibel sind. Die Menge der zu einem Basenpaar kompatiblen Basenpaare wird C-Menge dieses Basenpaares genannt. Im zweiten Verzweigungsschritt wird von den ersten Basenpaaren zu den Elementen ihrer C-Mengen verzweigt. Somit sind die Knotenpunkte der 2er Faltungen erreicht. Welche Basenpaare in eine Faltung aufgenommen sind, kann man feststellen, indem man vom Startpunkt bis zu dem Knotenpunkt im Graphen wandert, der die betreffende Faltung repräsentiert, und die Nummern der Basenpaare in allen durchlaufenen Knotenpunkten festhält. In allen weiteren Verzweigungsschritten wird dann jeweils nur noch zu denjenigen Basenpaaren verzweigt, die mit allen Basenpaaren kompatibel sind, die in den einzelnen Faltungen bereits aufgenommen sind.

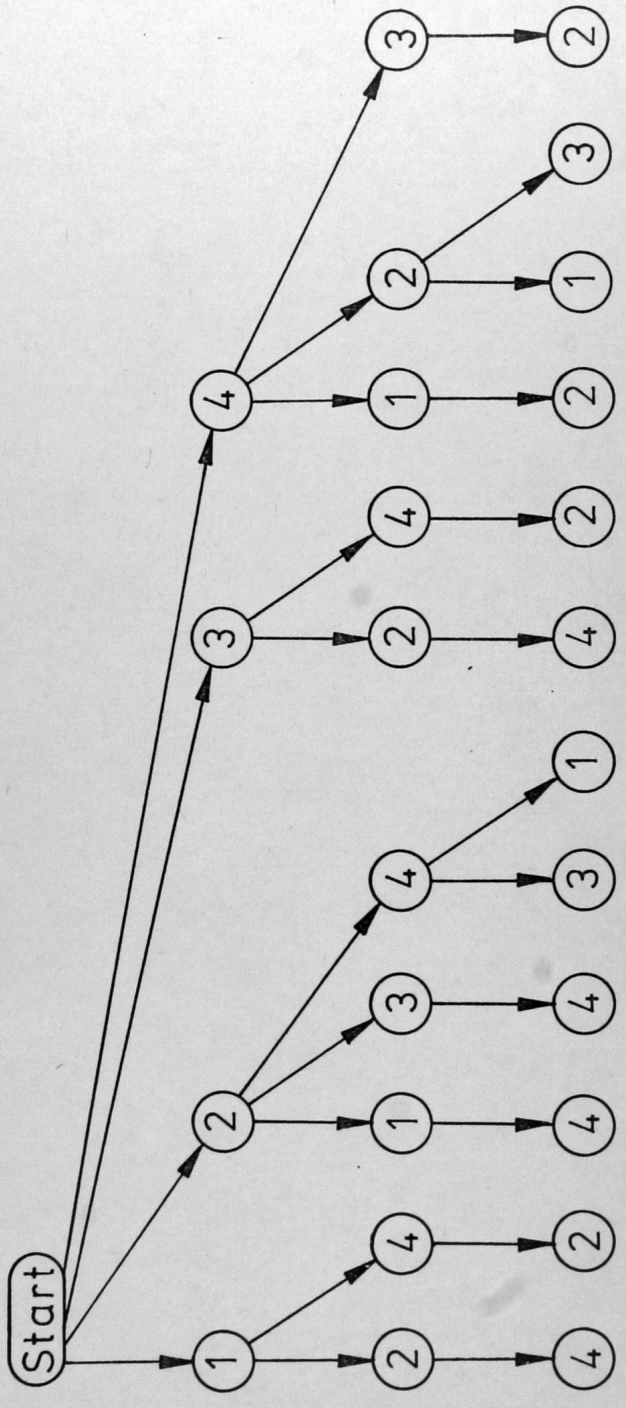


Abb.40:Graphische Darstellung des Zusammenhanges aller möglichen Faltungen mit vier Basenpaaren. 1.Stufe:1er Faltungen, 2. Stufe: 2er Faltungen,3.Stufe: 3er Faltungen.In einer Faltung sind alle Basenpaare vom Start bis zu dem betreffenden Knotenpunkt.

Man kann auch sagen, dass vom Knotenpunkt einer Faltung aus nur noch zu den Elementen der C-Menge der Faltung verzweigt wird. Hierzu sei der Begriff C-Menge noch einmal exakt definiert:

$$C((N\cdot N)_m) = \left\{(N\cdot N) \mid (N\cdot N)==(N\cdot N)_m\right\}$$

Eine Faltung ist eine Menge von Basenpaaren und es lässt sich entsprechend definieren: Es sei

$$C((N\cdot N)_m,(N\cdot N)_n) = \left\{(N\cdot N) \mid (N\cdot N)==(N\cdot N)_m \wedge (N\cdot N)==(N\cdot N)_n\right\}$$

für eine Faltung aus 2 Basenpaaren und allgemein

$$C((N\cdot N)_{m_1},(N\cdot N)_{m_2},\ldots,(N\cdot N)_{m_i}) = \left\{(N\cdot N) \mid \bigwedge_{n=1}^{i} (N\cdot N)==(N\cdot N)_{m_n}\right\}$$

für eine Faltung aus i verschiedenen Basenpaaren. Wie sich leicht nachweisen lässt, ist jede C-Menge einer Faltung mit i Basenpaaren Teilmenge der C-Menge der "Stammfaltung" dieser Faltung mit i-1 Basenpaaren. Dies gilt nicht, wenn die Faltung mit i-1 Basenpaaren in einem anderen Ast des Graphen vorkommt.

So ist in der Abbildung Nummer 40 das Basenpaar 2 kompatibel zu den Basenpaaren 1,3 und 4, Basenpaar 2 hat also die C-Menge 1,3,4 :

$$C((N\cdot N)_2) = \left\{(N\cdot N)_1,(N\cdot N)_3,(N\cdot N)_4\right\}$$

Geht man im Graphen vom Basenpaar Nummer 2 zum Basenpaar Nummer 4, so sind noch die Basenpaare 3 und 1 weiterhin kompatibel:

$$C((N\cdot N)_2,(N\cdot N)_4) = \left\{(N\cdot N)_3,(N\cdot N)_1\right\}$$

3 und 1 schliessen sich aber gegenseitig aus und der Graph besitzt an dieser Stelle keine weiteren Verzweigungen. Aus dem Graph lassen sich nun leicht alle möglichen Faltungen ablesen:

Faltungen mit 1 Basenpaar : 1; 2; 3; 4;
Faltungen mit 2 Basenpaaren: 1,2; 1,4; 2,1; 2,3;
2,4; 3,2; 3,4; 4,1;
4,2; 4,3;
Faltungen mit 3 Basenpaaren: 1,2,4; 1,4,2; 2,1,4;
2,3,4; 2,4,3; 2,4,1;
3,2,4; 3,4,2; 4,1,2;
4,2,1; 4,2,3; 4,3,2;

Faltungen mit mehr Basenpaaren sind nicht möglich. Dies sind insgesamt 26 verschiedene Faltungen. Nach unserer obigen Abschätzung sollte diese Anzahl kleiner als $4 + 4 \cdot 3 + 4 \cdot 3 \cdot 2 + 4 \cdot 3 \cdot 2 \cdot 1 = 64$ sein. Allein die Einführung eines einzigen inkompatiblen Basenpaares in unserem Beispiel hat die Anzahl der möglichen Faltungen um mehr als die Hälfte reduziert.

4.4.3 Möglichkeiten zur Verkürzung des Verfahrens

Nur, wenn sich geeignete Methoden zur Verkürzung des bisherigen Suchalgorithmus finden lassen, wird sich dieser ökonomisch zur völligen Optimierung von RNS-Sekundärstrukturen einsetzen lassen. In dieser Arbeit werden vier verschiedene Methoden verwendet: 1. Es wird keine Rücksicht auf die Reihenfolge der Faltung genommen. Von Faltungen, die die gleichen Basenpaare enthalten aber in anderer Reihenfolge wird nur eine bearbeitet. 2.Bindungsgleiche Basenpaare werden in einem Stück aufgenommen. 3.Es wird eine Abschätzung der freien Energie aller Faltungen eines Astes des Graphen vorgenommen anhand der Basenpaare der C-Menge. 4. Es wird eine Abschätzung mit Hilfe der 2er Sequenzen in den ungebundenen Bereichen erarbeitet. Mit Hilfe dieser Stabilitätsabschätzungen können unvorteilhafte Faltungen weggelassen werden.

4.4.3.1 Keine Berücksichtigung der Faltungsreihenfolge

Aus unserem letzten Beispiel kann man ersehen, dass die Faltungen 1,2; und 2,1; identisch sind, es ist unerheblich, welches Basenpaar als erstes gefaltet wurde, auch ist

$$C((N{\cdot}N)_1,(N{\cdot}N)_2) = C((N{\cdot}N)_2,(N{\cdot}N)_1) ,$$

das heisst, dass auch alle von der Faltung 1,2; ausgehenden Äste des Graphen sind gleich denen die von der Faltung 2,1; ausgehen, da die Menge der Basenpaare, die mit der einen Faltung vereinbar sind, gleich der Menge ist, die mit der anderen Faltung vereinbar sind. Es ist also nur nötig, entweder den Ast 1,2; oder den Ast 2,1; des Graphen zu bearbeiten. Willkürlich wurde festgelegt, dass immer nur die erste derartiger Alternativen zu bearbeiten ist, hat der Algorithmus 1,2; erreicht, überspringt er 2,1; . Durch diese Regel wird die Anzahl al-

ler möglichen Faltungen erneut verringert. In Abbildung 41 ist der nunmehr reduzierte Graph für unser Beispiel wiedergegeben:

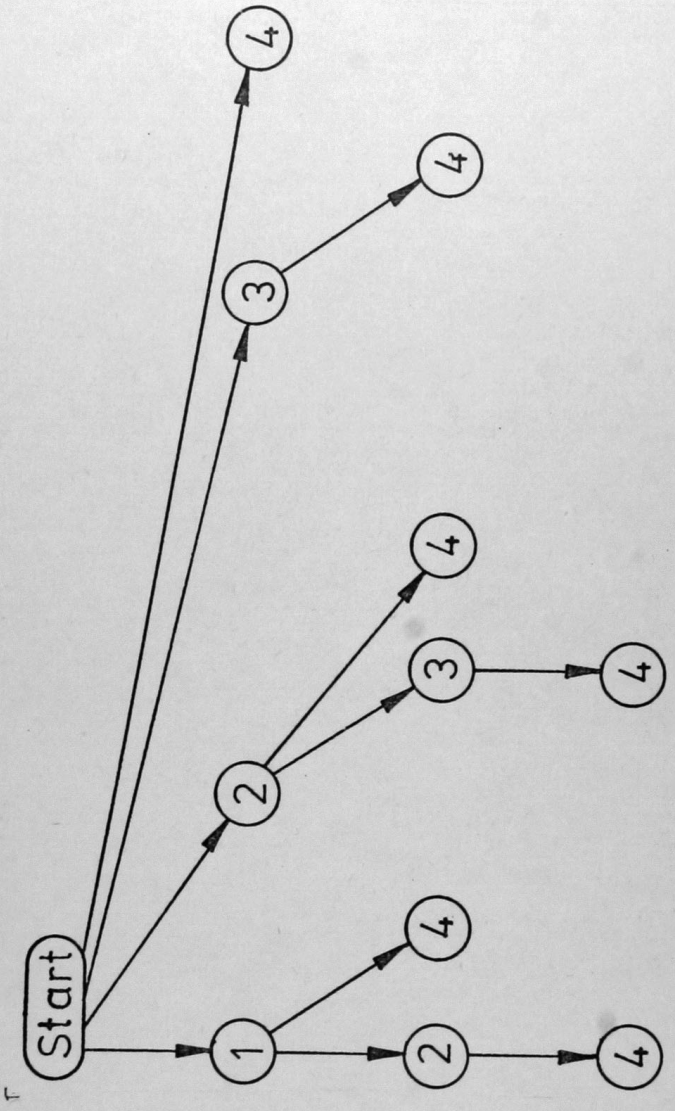


Abb.41: Reduzierter Graph aller möglichen Faltungen. Keine Berücksichtigung der Faltungsreihenfolge mehr.

Hieraus kann man entnehmen, dass es nunmehr nur noch 11 mögliche verschiedene Faltungen gibt:

1; 2; 3; 4; 1,2; 1,4; 2,3; 2,4; 3,4; 1,2,4; und 2,3,4; .

Kurz zusammengefasst bedeutet diese Regel, dass eine Faltung, die mit einem Basenpaar der Nummer m erweitert worden ist, nur noch mit Basenpaaren, deren Mummern grösser als m sind, erweitert zu werden braucht. Natürlich müssen diese Basenpaare in der C-Menge der betreffenden Faltung enthalten sein. Im Graphen sieht dies

so aus, dass eine Faltung nur noch mit Basenpaaren erweitert wird, die in Verzweigungen desselben Astes auftreten und rechts von ihr selbst liegen.

4.4.3.2 Bindungsgleiche Basenpaare

Es kommt vor, dass eine Faltung durch 2 verschiedene Basenpaare erweitert werden kann, die miteinander kompatibel sind und zu der gleichen Reduktion der C-Menge führen, d.h. dass

$\{(N\cdot N)_m\} + C((N\cdot N)_m) = \{(N\cdot N)_n\} + C((N\cdot N)_n)$ ist.

Man kann jede Faltung, die das Basenpaar m enthält mit n erweitern, ohne dass die C-Menge der Faltung weiter verringert wird. Es ist darum sinnvoll diese beiden Basenpaare in einem Schritt in eine Faltung aufzunehmen, a1/ so als "gekoppelt" zu betrachten. Solche Basenpaare werden in der Folge als "bindungsgleiche Basenpaare" bezeichnet. Es handelt sich hierbei vorwiegend um Basenpaare, die in ein und derselben Helix vereinigt sind, die sich mit keiner weiteren Helix in der Bindungsmatrix nochmal überschneidet. Dies kommt vorwiegend gegen Ende eines Faltungsprozesses vor, wenn nur noch wenige Helices oder Helixteile in den C-Mengen enthalten sind, oder wenn es sich um Faltungen von kleinen RNS-Abschnitten handelt. Gruppiert man die Basenpaare in den C-Mengen derart, dass diejenigen, die in eine Helix gehören, nebeneinander stehen, so lässt sich in den Suchalgorithmus leicht eine Abfrage auf Bindungsgleichheit einbauen. Hierdurch werden die Äste in der Darstellung des Graphen zwar nicht weniger aber kürzer. Siehe Abbildung 42.

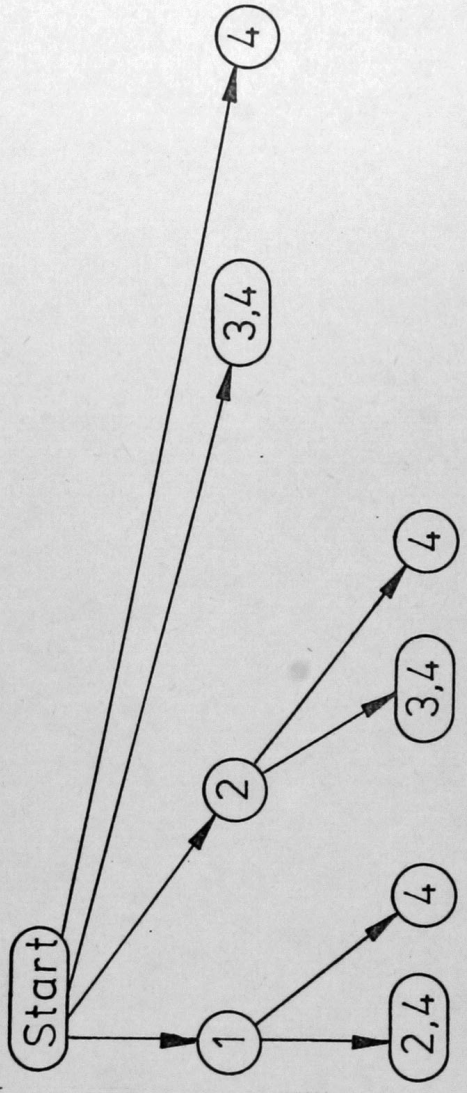


Abb. 42: Bindungsgleiche Basenpaare wurden zu einem einzigen Knotenpunkt zusammengefasst und die Äste des Graphen somit verkürzt.

4.4.3.3 Abschätzung der freien Energie der C-Menge

Die beiden letzten Verfahren zur Verkürzung des Suchalgorithmus basieren auf einer Abschätzung der thermodynamischen Stabilität, die durch Faltungen in einem bestimmten Ast des Graphen bestenfalls erreicht werden kann. Liegt eine solche Abschätzung oberhalb der freien Energie einer bekannten Faltung desselben RNS-Stückes, so kann auf eine weitere Bearbeitung dieses Astes verzichtet werden. Die Errechnung einer solchen Abschätzung muss einfacher sein, als die Bearbeitung des Astes selbst.

Als erster Schritt der Abschätzung wird die Gesamt-freie Energie der bereits erhaltenen Faltung des betreffenden Astes errechnet. Hierzu werden die negativen freien Energien aller in der C-Menge durch Basenpaare vertretenen Helices und Teilhelices hinzugezählt, einschliesslich der kleinsten möglichen negativen freien Energien aller noch denkbaren bulge loops (d.h. der Stapelkräfte, die zwischen den beiden Basenpaaren an Anfang und Ende eines bulges auftreten). Da die positive freie Energie des Gesamtmoleküls bei Einlagerung einer weiteren Helix nur zunehmen kann, ist diese Summe der niedrigste Gesamtfreie Energie-Wert, den keine Faltung, die innerhalb dieses Astes noch gebildet werden kann, zu unterschreiten in der Lage ist.

4.4.3.4 Abschätzung mit Hilfe von 2er Sequenzen

Weiss man, dass in einer Faltung in einem ungebundenen Teil zweimal die 2er Sequenz A-U auftritt, so lässt sich nicht ausschliessen, dass die Faltung durch $\begin{matrix}-A-U-\\ \cdot\ \cdot\\ -U-A-\end{matrix}$ erweitert werden kann. Zwei A-U-Sequenzen können in Basenpaarbindung zueinander treten, vorausgesetzt, sie liegen nicht zu nah beieinander oder liegen in Teilen des Moleküls, zwischen denen eine Basenpaarbindung ausgeschlossen ist. Auch können zwei verschiedene Sequenzen, wie z.B. A-A und U-U sich miteinander koppeln. Es brauchen für eine Abschätzung nur solche Kopplungen betrachtet zu werden, deren freie Energie kleiner als 0.0 ist. Alle "selbstbindenden" 2er Sequenzen sind:

A-U,U-A,G-C,C-G,U-G,G-U

Die paarweise koppelnden Sequenzen, die berücksichtigt werden sind:

A-A/U-U, A-G/C-U, A-C/G-U, U-G/C-A, U-C/G-A, G-G/C-C

Durch Auszählung der verschiedenen 2er Sequenzen und Errechnung der möglichen Anzahl von Kopplungen jeder Art erhält man eine Abschätzung der maximal noch möglichen 2er Helices bei der betreffenden Faltung. Dies muss natürlich getrennt für jede Schleife extra geschehen, weil die Nucleotide innerhalb einer Schleife sich nicht mit Nucleotiden ausserhalb binden können. Die Summe der freien Energien aller noch möglichen 2er Helices wird zu der errechneten Gesamt-freien Energie der betreffenden Faltung hinzugerechnet. Auch hierbei müssen die möglichen negativen freien Energien von bulge loops berücksichtigt werden. Diese Summe gibt eine weitere Schätzung der für einen Ast des Graphen im besten Falle erreichbaren Gesamtfreien Energie und somit die Aussicht auf diesem Wege eine Faltung aufzufinden, die stabiler ist als eine bereits bekannte.

4.4.4 Realisierung durch das Programm MODELL

In Abbildung 43 ist der Weg des Programms MODELL (gestrichelte Linie) eingezeichnet, den es verfolgt, wenn es den Graph, der die möglichen Faltungen darstellt, absucht, wiederum dargestellt an dem Beispiel der 4 Basenpaare aus Abschnitt 4.4.2 . Es wird soweit wie möglich senkrecht fortgeschritten, bis eine maximale Faltung erreicht worden ist, dann wird der letzte Schritt zu dieser Faltung wieder rückgängig gemacht und weitergegangen. Es wird sodann das nächste Element der C-Menge in die Faltung aufgenommen und wieder soweit wie möglich senkrecht fortgeschritten, usw., bis alle Faltungen gefunden und getestet worden sind. Die einzelnen Äste des Graphen werden also von links nach rechts der Reihe nach abgesucht. Ordnet man die Basenpaare im ersten Verzweigungsschritt nach dem Start so, dass links diejenigen liegen, die den stabilsten Helices angehören und rechts diejenigen, die Teile von geringer stabilen Helices sind, so werden durch diesen Algorithmus die vermutlich stabilsten Faltungen als erste generiert. Gegen Ende des Programms sind dann auch die Abschätzungen der prospektiven Stabilität des betreffenden Astes häufiger von Erfolg.

Um im Computer Speicherplatz einzusparen, wird niemals der vollständige Graph in der Rechenmaschine datenmässig repräsentiert. Während der Suche nach Faltungen kennt das Programm nur die Knotenpunkte, die in gerader Linie vom Startpunkt bis zu dem gerade bearbeiteten Knotenpunkt liegen. Die Daten, die den restlichen Knotenpunkten entsprechen, wie z.B. die Informationen über die an den Knoten aufzunehmenden Basenpaare oder die C-Menge derselben, sind entweder schon bearbeitet und "vergessen", d.h. der gleiche Speicherplatz kann erneut ver-

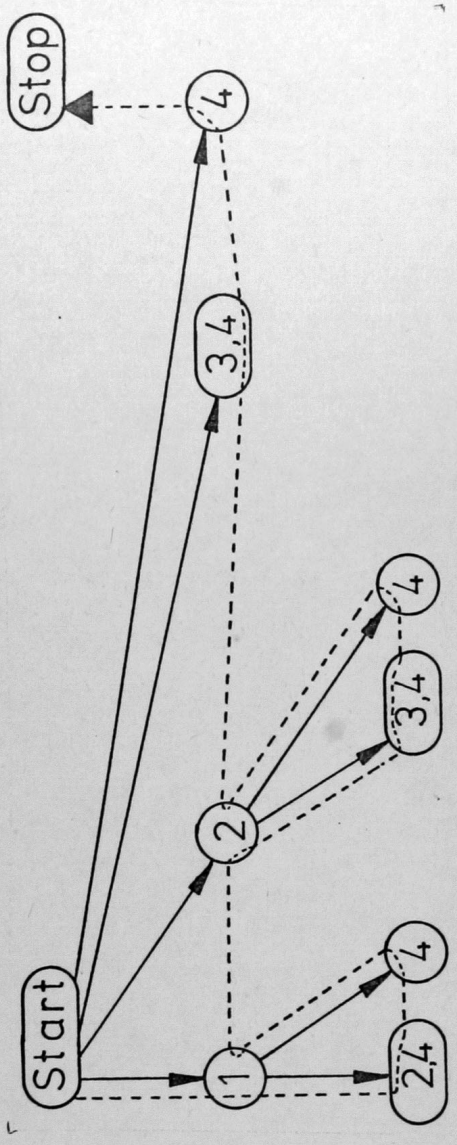


Abb. 43: Weg, den das Programm MODELL verfolgt, wenn es einen Graphen, der die Menge der möglichen Faltungen darstellt, absucht.

wendet werden, oder noch nicht bekannt, sofern es sich um Knoten handelt, die vom Algorithmus noch nicht erreicht worden sind. Durch die Anpassungsfähigkeit der Programmiersprache SIMULA ist es möglich, während des gesamten Programmverlaufs, benötigten Speicherplatz neu anzufordern und überflüssige Teile wieder freizugeben. Anders wäre ein solches Programm nicht durchführbar. In Abbildung 44 ist der genaue Algorithmus dargestellt, wie er in MODELL inkorporiert ist.

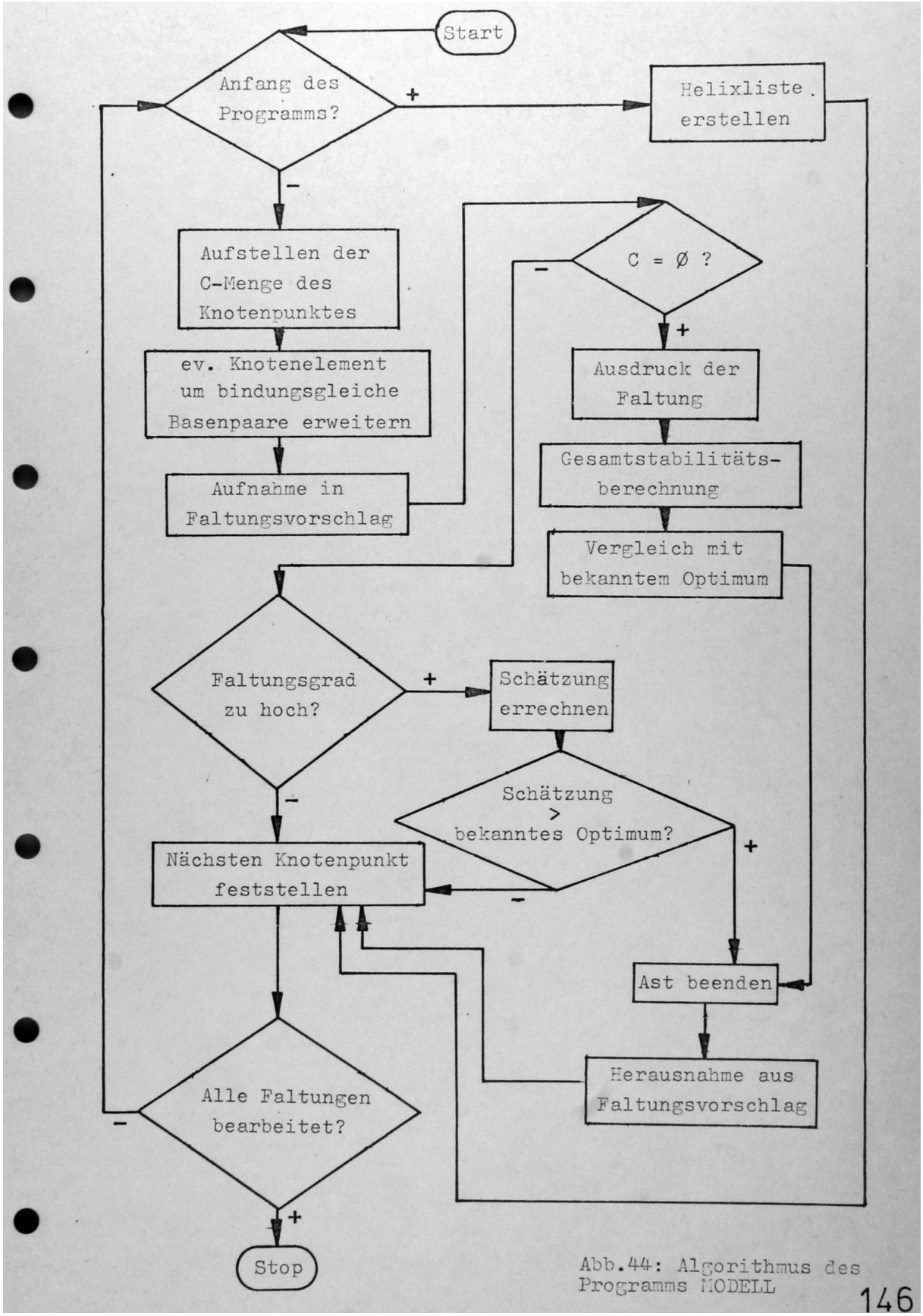


Abb.44: Algorithmus des Programms MODELL

Grosse Teile des Programms konnten den Programmen HELIX.LISTE und SIM.FALTUNG entnommen werden. Vornehmlich handelte es sich um die Erstellung der Helixliste, Begrenzung der Helices, Ausdruck der Faltungen und Berechnung der Gesamt-freien Energie. Das Programm MODELL hat eine Länge von 1172 Lochkarten und besitzt 44 Unterprogramme.

4.4.5 Ergebnisse der völligen Optimierung

Das Programm MODELL für die völlige Optimierung von RNS-Sekundärstrukturen ist erst sehr kurzfristig vor Abschluss der Diplomarbeit fertig geworden. So war es nur möglich quasi in letzter Minute drei kürzere Beispielsequenzen bearbeiten zu lassen. Es handelt sich hierbei um die Teilsequenzen 1-20, 1-40 und 1-60 von MS2 , Sequenzen also, die in einem langsamen "scale up" um 20 Nucleotide jeweils verlängert wurden, um festzustellen bis zu welcher Länge das Programm eine Nucleinsäure ökonomisch bearbeiten kann.
Die Teilsequenz 1-20 wurde vollständig bearbeitet und das völlige Optimum aufgefunden, es stimmt mit dem einfachen Optimum überein und ist dieselbe Sekundärstruktur die in Abbildung 16 für diese Teilsequenz wiedergegeben ist.
Um abzusichern, dass es sich bei dieser Faltung auch tatsächlich um das völlige Optimum handelt, also um diejenige Sekundärstruktur mit der niedrigsten Gesamt-freien Energie, wurden von MODELL alle weiteren 625 verschiedenen möglichen Sekundärstrukturen berechnet, bis zum Ende des Suchalgorithmus. Von jeder möglichen Faltung wurde die Gesamt-freie Energie berechnet und eine Liste der Häufigkeiten von Faltungen mit bestimmter Stabilität angefertigt:

Liste der Anzahlen der Faltungen des Abschnittes 1-20 von MS2 geordnet nach der freien Energie:

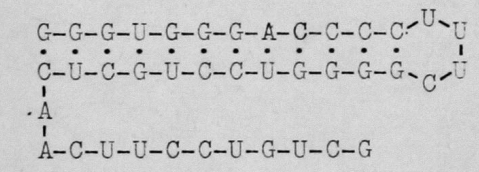
Freie Energie(kJ)	Anzahl
-31.39125 <= ΔG < -27.20575	1
-14.64925 <= ΔG < -10.46375	1
-10.46375 <= ΔG < -6.27825	2
-6.27825 <= ΔG < -2.09275	2
-2.09275 <= ΔG < 2.09275	4
2.09275 <= ΔG < 6.27825	14
6.27825 <= ΔG < 10.46375	5
10.46375 <= ΔG < 14.64925	9

Freie Energie (kJ)	Anzahl
14.64925 <= ΔG < 18.83475	16
18.83475 <= ΔG < 23.02025	28
23.02025 <= ΔG < 27.20575	21
27.20575 <= ΔG < 31.39125	24
31.39125 <= ΔG < 35.57675	35
35.57675 <= ΔG < 39.76225	61
39.76225 <= ΔG < 43.94775	59
43.94775 <= ΔG < 48.13325	71
48.13325 <= ΔG < 52.31875	69
52.31875 <= ΔG < 56.50425	65
56.50425 <= ΔG < 60.68975	37
60.68975 <= ΔG < 64.87525	29
64.87525 <= ΔG < 69.06075	32
69.06075 <= ΔG < 73.24625	17
73.24625 <= ΔG < 77.43175	14
77.43175 <= ΔG < 81.61725	7
81.61725 <= ΔG < 85.80275	3

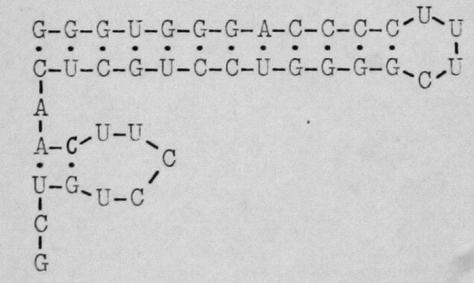
Nur 6 Faltungen haben eine freie Energie, die eindeutig kleiner als 0.0 kJ ist (< -2.09 kJ). Alle anderen Faltungen müssen als thermodynamisch unvorteilhaft angesehen werden, weil zu ihrer Bildung Energie aufgewendet werden müsste.

Die Optimierung der Teilsequenz 1-40 konnte systembedingt nicht vollständig durchgeführt werden. Der Algorithmus stellte 15002 verschiedene Faltungen auf. Gleich als dritte dieser Faltungen wurde eine Struktur gefunden, deren freie Energie auf -69.89785 kJ berechnet wurde. Dieser Wert wurde von keiner der weiteren 14999 Faltungen mehr unterschritten, so dass sich vermuten lässt, dass es sich hier um das völlige Optimum handelt. Durch die Anordnung der Basenpaare im ersten Verzweigungsschritt der Graphendarstellung der Menge aller möglichen Faltungen wird dafür gesorgt, dass die vermutlich stabilsten Faltungen als erste generiert werden. So wurde in den letzten 500 generierten Faltungen keine Faltung mehr gefunden mit einer Gesamt-freien Energie kleiner

als -61.0 kJ. Überhaupt sind die meisten der Faltungen gegen Ende des Programmlaufs energetisch ungünstig, haben also positive freie Energie-Werte. Die völlige Optimierung wurde also nicht bis zum Ende durchgeführt, gelangte aber bis zu einem Punkt im Graphen der möglichen Faltungen, ab dem das Auffinden weiterer noch stabilerer Faltungen sehr unwahrscheinlich ist. Für die Berechnung der 15002 Faltungen benötigte das Programm 1 Minute und 47 Sekunden. Eine völlige Optimierung von RNS-Abschnitten der Länge 40 ist also innerhalb tragbarer Rechenzeiten ohne weitere Verkürzung des Suchalgorithmus möglich. Das vermutliche völlige Optimum der Teilsequenz 1-40 von MS2 stimmt nicht mit dem einfachen Optimum überein, das für dieselbe Sequenz von dem Programm FALTUNG bzw. SIM.FALTUNG erarbeitet wurde. Das einfache Optimum besitzt nur eine freie Energie von -53.99295 kJ; liegt also deutlich über dem Wert -69.89785 kJ der von MODELL aufgefundenen Faltung. Die Faltung nach MODELL hat folgende Struktur:



Hingegen besitzt das einfache Optimum nach FALTUNG und SIM.FALTUNG eine Helix mehr:



Das Programm MODELL ist also in der Lage, durch Ausschluss energetisch ungünstiger Helices eine optimalere Struktur zu finden als die Programme der einfachen Optimierung.

In der folgenden Tabelle ist die Verteilung der freien Energien aller untersuchten Faltungen des Abschnittes 1-40 wiedergegeben:

Liste der Anzahl der Faltungen des Abschnittes 1-40 von MS2 geordnet nach der freien Energie

Freie Energie (kJ)			Anzahl
-73.24625	<= Δ G <	-69.06075	1
-60.68975	<= Δ G <	-56.50425	3
-56.50425	<= Δ G <	-52.31875	1
-52.31875	<= Δ G <	-48.13325	1
-48.13325	<= Δ G <	-43.94775	7
-43.94775	<= Δ G <	-39.76225	1
-39.76225	<= Δ G <	-35.57675	6
-35.57675	<= Δ G <	-31.39125	17
-31.39125	<= Δ G <	-27.20575	13
-27.20575	<= Δ G <	-23.02025	37
-23.02025	<= Δ G <	-18.83475	35
-18.83475	<= Δ G <	-14.64925	31
-14.64925	<= Δ G <	-10.46375	112
-10.46375	<= Δ G <	-6.27825	54
-6.27825	<= Δ G <	-2.09275	95
-2.09275	<= Δ G <	2.09275	190
2.09275	<= Δ G <	6.27825	167
6.27825	<= Δ G <	10.46375	248
10.46375	<= Δ G <	14.64925	286
14.64925	<= Δ G <	18.83475	349
18.83475	<= Δ G <	23.02025	425
23.02025	<= Δ G <	27.20575	454
27.20575	<= Δ G <	31.39125	570
31.39125	<= Δ G <	35.57675	529
35.57675	<= Δ G <	39.76225	691
39.76225	<= Δ G <	43.94775	598
43.94775	<= Δ G <	48.13325	776
48.13325	<= Δ G <	52.31875	670
52.31875	<= Δ G <	56.50425	779
56.50425	<= Δ G <	60.68975	790
60.68975	<= Δ G <	64.87525	695
64.87525	<= Δ G <	69.06075	771
69.06075	<= Δ G <	73.24625	727
73.24625	<= Δ G <	77.43175	731

Freie Energie (kJ)		Anzahl
77,43175 <= ΔG <	81.61725	681
81.61725 <= ΔG <	85.80275	570
85.80275 <= ΔG <	89.98825	640
89.98825 <= ΔG <	94.17375	423
94.17375 <= ΔG <	98.35925	487
98.35925 <= ΔG <	102.54475	369
102.54475 <= ΔG <	106.73025	301
106.73025 <= ΔG <	110.91575	201
110.91575 <= ΔG <	115.10125	150
115.10125 <= ΔG <	119.28675	172
119.28675 <= ΔG <	123.47225	58
123.47225 <= ΔG <	127.65775	46
127.65775 <= ΔG <	131.84325	28
131.84325 <= ΔG <	136.02875	11
136.02875 <= ΔG <	140.21425	2
140.21425 <= ΔG <	144.39975	3

Von 15002 Faltungen haben also nur 414 eine freie Energie, die sicher unter 0.0 liegt.

Die Suche nach der stabilsten Konformation des Abschnittes 1-60 von MS2 wurde nach 402 Faltungen abgebrochen. Zu ihrer Berechnung benötigte das Programm 12 Sekunden. Es erstellt die Faltungen also etwa 4 mal langsamer als bei der Sequenz 1-40 . Aber bereits unter diesen 402 Faltungen wurde eine Faltung aufgefunden deren Gesamt-freie Energie gleich -74.92045 kJ war . Durch einfache Optimierung konnte nur eine Faltung mit -41.43645 kJ gefunden werden,sodass hier fast eine Verdopplung der Stabilität eingetreten ist. Auf der nächsten Seite werden diese beiden Faltungen -rechts die durch MODELL aufgefundene und links die einfach optimierte- in linearer Form wiedergegeben, so wie sie direkt vom Programm ausgedruckt werden. Es stehen hierbei neben den Nucleotiden deren Nummern. Die Numerierung der Nucleotide ist wie bereits erwähnt vom 5'-Ende anfangend durchgehend bis zum 3'-Ende.

1	G•C	28	1	G•C	28
2	G•U	27	2	G•U	27
3	G•C	26	3	G•C	26
4	U•G	25	4	U•G	25
5	G•U	24	5	G•U	24
6	G•C	23	6	G•C	23
7	G•C	22	7	G•C	22
8	A•U	21	8	A•U	21
9	C•G	20	9	C•G	20
10	C•G	19	10	C•G	19
11	C•G	18	11	C•G	18
12	C•G	17	12	C•G	17
13	U		13	U	
14	U		14	U	
15	U		15	U	
16	C		16	C	
17	G•C	12	17	G•C	12
18	G•C	11	18	G•C	11
19	G•C	10	19	G•C	10
20	G•C	9	20	G•C	9
21	U•A	8	21	U•A	8
22	C•G	7	22	C•G	7
23	C•G	6	23	C•G	6
24	U•G	5	24	U•G	5
25	G•U	4	25	G•U	4
26	C•G	3	26	C•G	3
27	U•G	2	27	U•G	2
28	C•G	1	28	C•G	1
29	A		29	A•U	56
30	A•U	38	30	A	
31	C•G	37	31	C	
32	U		32	U	
33	U		33	U	
34	C		34	C	
35	C		35	C	
36	U		36	U•A	51
37	G•C	31	37	G•C	50
38	U•A	30	38	U	
39	C		39	C	
40	G		40	G	
41	A		41	A	
42	G•C	49	42	G•C	49
43	C•G	48	43	C•G	48
44	U		44	U	
45	A		45	A	
46	A		46	A	
47	U		47	U	
48	G•C		48	G•C	43
49	C•G		49	C•G	42
50	C•G	60	50	C•G	37
51	A•U	59	51	A•U	36
52	U•A	58	52	U	
53	U		53	U	
54	U		54	U	
55	U		55	U	
56	U		56	U•A	29
57	A		57	A	
58	A•U	52	58	A	
59	U•A	51	59	U	
60	G•C	50	60	G	

Links: einfach optimiert

Rechts: aufgefunden von MODELL

Die lineare Schreibweise hat den Vorteil, dass man an ihr sofort sehen kann welche Basenpaare zweier Faltungen gleich sind und welche nicht. Die Faltung nach MODELL gewinnt ihre Stabilität gegenüber der einfach optimierten durch eine erhöhte Bildung von bulge loops und internal loops, d.h. dass mehr Helices sich im Haarnadelbereich einer anderen Helix gebildet haben als bei der Vergleichsfaltung. Es ist aber von MODELL in diesen 402 ersten Faltungen sicher noch nicht das völlige Optimum erreicht worden, dies ersieht man allein aus der Präsenz des einzelnen Basenpaares ($U_{56} \cdot A_{29}$), das keinen Beitrag zur Stabilität leistet (Einzelne Basenpaare in einer Faltung erhalten eine freie Energie von 0.0 kJ!).Aber allein durch diese kurze Suche nach stabileren Faltungen konnte eine um-33.484 kJ stabilere Faltung aufgefunden werden als nach dem bisherigen Verfahren.

Liste der Anzahlen der Faltungen des Abschnittes 1-60 von MS2 geordnet nach der freien Energie

Freie Energie	Anzahl
-77.43175 <= ΔG < -73.24625	3
-73.24625 <= ΔG < -69.06075	1
-69.06075 <= ΔG < -64.87525	8
-64.87525 <= ΔG < -60.68975	4
-60.68975 <= ΔG < -56.50425	6
-56.50425 <= ΔG < -52.31875	21
-52.31875 <= ΔG < -48.13325	23
-48.13325 <= ΔG < -43.94775	23
-43.94775 <= ΔG < -39.76225	31
-39.76225 <= ΔG < -35.57675	48
-35.57675 <= ΔG < -31.39125	44
-31.39125 <= ΔG < -27.20575	39
-27.20575 <= ΔG < -23.02025	36
-23.02025 <= ΔG < -18.83475	29
-18.83475 <= ΔG < -14.64925	23
-14.64925 <= ΔG < -10.46375	25
-10.46375 <= ΔG < -6.27825	15
-6.27825 <= ΔG < -2.09275	19
-2.09275 <= ΔG < 2.09275	5
2.09275 <= ΔG < 6.27825	1

Es ergab sich, dass nur 6 der aufgefundenen Faltungen eine freie Energie hatten, die nicht sicher unter 0.0 kJ lag.
Je länger also ein von Modell bearbeitetes RNS-Stück wird, desto mehr erhöht sich der Anteil derjenigen Faltungen, die eine freie Energie kleiner als 0.0 kJ haben, also thermodynamisch möglich sind. Je länger allerdings das Programm arbeitet, desto mehr unvorteilhafte Faltungen produziert es.

Es wäre wahrscheinlich möglich gewesen von Hand leichter zu einer optimalen Faltung von den Teilsequenzen 1-20, 1-40 und 1-60 zu gelangen, als durch Aufstellung des umfangreichen Programms MODELL. Hier sollte jedoch nachgewiesen werden, dass es möglich ist, algorithmisch zu einer solchen optimalen Faltung zu gelangen. Jetzt, nachdem die ersten Läufe des Programms durchgeführt sind, werden weitere Methoden und Wege sichtbar, mit denen man zu Abkürzungen des Programms kommen kann. Es liessen sich z.B. Wahrscheinlichkeiten ausrechnen, für das Auftreten von stabilen Faltungen in den einzelnen Ästen des Graphen. Es würde u.U. genügen, wenn man nur Äste betrachtet, die eine derartige "prospektive" Wahrscheinlichkeit grösser als 0.001 hätten oder ähnlich. Man könnte dann von dem erhaltenen Optimum nur noch behaupten, dass es mit einer bestimmten Wahrscheinlichkeit p(Optimum) auch die gesuchte Faltung ist. Wenn diese Wahrscheinlichkeiten aber niedrig genug zu halten wären, würden solche Faltungen mit genau angebbarer Sicherheit die optimalen sein.
MODELL ist also keineswegs ein abgeschlossenes Verfahren, sondern ein Anfangspunkt für die Aufstellung von noch reduzierteren Algorithmen, die in der Lage sein könnten, RNS Sequenzen von beträchtlicher Länge zu falten und nicht nur die hier betrachteten Teilstücke.

5 Diskussion

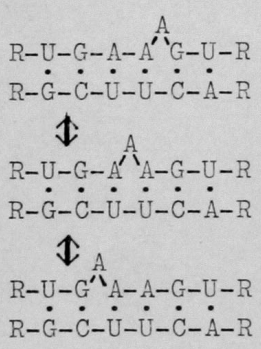
Es hat sich gezeigt, dass es möglich ist, sichere Aussagen über die Sekundärstruktur einer RNS zu machen, ohne einen konkreten Vorschlag einer bestimmten Faltung vorzulegen, ja selbst ohne die genaue Primärsequenz des Moleküles zu kennen. Man kann die Kenntnis der Basenzusammensetzung, eine nearest-neighbour-Analyse und eine Zählung der möglichen helicalen Bereiche heranziehen, um über die Fähigkeit der RNS zur Bildung von Sekundärstrukturen Aussagen machen zu können. MS2 und ΦX174 zeigen deutlich bessere Faltungskapazitäten als vergleichbare RNS-Moleküle mit gleichmässiger Basenverteilung und zufälliger Sequenz. Die Informationskapazitäten der Bakteriophagen haben dadurch nur unwesentlich gelitten. Allerdings wäre eine erneute Berechnung der Informationseinschränkungen durch das beobachtete Ausmass an 3er und 10er Helices bei MS2 sinnvoll, da bei der bisherigen Berechnung die statistischen Abhängigkeiten zwischen den Helices zum grossen Teil ausser acht gelassen wurden und es sich nicht sagen lässt, inwieweit diese das Ergebnis verändern.

Weiterhin zeigt der Vergleich der einfach optimierten Faltungen von MS2 und einem Zufallsmessenger, dass von MS2 stabilere Faltungen mit höherem Bindungsgrad erstellt werden können, als von beliebigen anderen RNS.

Aber immer noch haben die per Computer generierten Faltungen geringere Stabilitätswerte als die in der Literatur wiedergegebenen, die z.T. auf experimentellen Daten beruhen. Es scheint somit notwendig, die bestehenden Faltungsalgorithmen weiter zu verbessern. Eine eindeutige oder völlige Optimierung ist nur für kurzkettige RNS-Moleküle gangbar, für längere RNS müssen approximative Verfahren gesucht werden, d.h. Verbesserungen der einfachen Optimierung.

Die Computer-Gesamtfaltung von MS2 hat zwar eine geringere thermodynamische Stabilität als der Sekundärstrukturvorschlag der Arbeitsgruppe FIERS, es zeigen sich aber beträchtliche Übereinstimmungen hinsichtlich der verwendeten Helices. Die Computerfaltung beinhaltet nur etwa 0.1% aller möglichen Basenpaare, diese Basenpaare stimmen aber zu etwa 25% mit denjenigen überein, die nach FIERS in der Faltung vorkommen. Selbst die einfache Optimierung stellt also durchaus ein leistungsfähiges Verfahren dar.

Je länger eine Nucleotidkette ist, desto weniger sinnvoll ist es, nach einem eindeutigen Optimum seiner Sekundärstrukturen zu suchen. Es kann davon ausgegangen werden, dass sich auch in vivo lokale Eigenschaften des Moleküles in konstanter Veränderung befinden. Z.B. gibt es "wandernde Bulges", die alle gleichwertige Konformationen desselben Moleküls darstellen:



Thermodynamisch sind diese Konformationen nicht zu unterscheiden. Es ist darum besser nicht von <u>der</u> optimalen Konformation zu sprechen, sondern von einer Klasse von optimalen Konformationen, die im gegebenen Fall recht umfangreich sein kann. Deshalb kann man unter Umständen Konformationsvorschäge nicht falsifizieren, msn müsste stattdessen angeben, mit welcher Häufigkeit das Molekül in dieser oder jener Konformation vorliegt. In Fällen, wie dem

obigen Beispiel liegen alle Konformationen bei Ausschluss weiterer äusserer Faktoren mit gleicher Häufigkeit vor.
Wenn eine experimentell festgestellte Faltung nicht mit einer unter Verwendung der bekannten thermodynamischen Parameter optimierten Faltung übereinstimmt, liegt das u.U. nicht an dem verwendeten Optimierungsverfahren, sondern an der Ungenauigkeit der Parameter. Ein strenger Vergleich von optimierten und experimentell erhaltetenen Faltungen kann Hinweise geben, welche Parameter insbesondere verbessert werden müssen. Es ist vielleicht möglich, theoretisch festzustellen, durch welche Parameterveränderungen die optimierten Faltungen mehr den experimentellen entsprechen. Diese veränderten Parameter müssten dann ihrerseits experimentell überprüft werden. Auf diese Weise könnten sich Theorie und Praxis sinnvoll ergänzen.
Notwendig ist weiterhin ein Vergleich der Daten, die durch elektronenmikroskopische Untersuchung der Konformationen von RNS, speziell von MS2 vorliegen, und den möglichen Helices, wie sie sich in der Bindungsmatrix zeigen. MS2 wurde 1976 von JACOBSON (19) untersucht. Je nach der Ionenstärke des Mediums zeigten sich 1 bis 3 grosse offene Schleifen, d.h. Helices mit Haarnadelschleifen die 10 bis 20% des Moleküles umfassten, und Helices mit im Elektronenmikroskop nicht sichtbaren Schleifen, die etwa 3-5% der RNS-Länge ausmachten. Die mittlere Komplexität der Faltungen nahm mit der Salzkonzentration zu und es war möglich, ein sich wiederholendes Muster von Sekundärstrukturbildungen zu beobachten. Es liessen sich sicherlich exakt diejenigen Helices in der Bindungsmatrix lokalisieren, aus denen sich die genannten Schleifen zusammensetzen. Da sich weiterhin experimentell feststellen lässt, welche helicalen Bereich sich bei Erhöhung der Ionenstärke des Mediums als erste bilden, liessen sich daraus wertvolle Hinweise gewinnen, mit denen es vielleicht möglich wäre

den Algorithmus der einfachen Optimierung zu verbessern.

Ein nächster Schritt der Untersuchung der Konformation von Nucleinsäuren ist die Berücksichtigung der wahren atomaren Abstände, wie dies z.B. von der Arbeitsgruppe KIM (22,23) versucht wird. Es entstand im Kapitel 4.3.1 die Frage, ob die Basenpaarbindung zwischen den Nucleotiden innerhalb einer Schleife und ausserhalb möglich sei oder nicht. Da diese Frage im Rahmen dieser Arbeit nicht beantwortet werden konnte wurde festgesetzt, dass diese Bindungen von den weiteren Überlegungen auszuschliessen seien. Der wesentliche Grund für diese Festlegung war die Unmöglichkeit, mit den bisherigen Mitteln eine Frage zu beantworten, die nur unter Berücksichtigung der wahren atomaren Abstände gelöst werden kann. Ein Programm, das dieses leisten würde, könnte auch imstande sein, festzustellen inwieweit Interaktionen zwischen den einzelnen helicalen Bereichen bei der Bildung der Konformation des Moleküles eine Rolle spielen. So könnte man nicht nur eine Sekundärstruktur aufstellen, sondern auch eine Tertiär- und Quartärstruktur vorschlagen. Aufgrund der Vielzahl von Faktoren, die bei einer solchen Analyse zu berücksichtigen wären (Bindungsabstände, Bindungswinkel, Ladungen etc. für alle beteiligten Atome) ist ein solches Vorgehen für längerkettige Nucleinsäuren bislang noch nicht gangbar. Auch hier müssen Möglichkeiten erarbeitet werden, den Umfang des Problems zu verringern.

Obwohl also sowohl mit den Programmen, die allein auf der Abstraktionsebene der Nucleotidsequenz arbeiten, wie auch mit den Programmen, die die wahren atomaren Abstände berücksichtigen noch keine thermodynamisch optimalen Faltungen sehr langer RNS-Sequenzen erstellt werden können, kann man mit ihnen sehr wohl

begrenztere Fragestellungen bearbeiten. Man kann sich Vorstellungen über die Grobstruktur des Moleküles erarbeiten, ob es z.B. wahrscheinlich globulär oder gestreckt vorliegt, man kann lokale Helices mit sehr grosser Bildungswahrscheinlichkeit von nicht-lokalen unterscheiden, es ist möglich für einige Teile des Moleküls anzugeben, ob es sich wahrscheinlicher an der Aussenseite des Moleküls oder im Innern aufhält. Auch Hypothesen über wahrscheinliche physiologische Funktionen der vermuteten Sekundärstruktur lassen sich bilden. Andererseits werden für solche Moleküle u.U. Funktionen postuliert, denen es aufgrund seiner Sekundärstrukturgegebenheiten nicht gerecht werden kann, was aber ohne eine genauere Untersuchung nicht entscheidbar ist, wo also ein Programm der bisherigen Art wertvolle Dienste leisten könnte.

Liste aller durch ihre Basenpaare unterschiedenen Helices im Abschnitt 1 bis 100 von MS2

Programm: ZAEHLUNG ΔG in kcal

Anzahl	ΔG	Sequenz
1	0.0	-C-C-G- • • • -U-G-U-
1	0.0	-G-U-C- • • • -C-G-G-
1	0.0	-U-G-C- • • • -A-U-G-
3	0.0	-G-C-U- • • • -U-G-G-
2	0.0	-U-U-A- • • • -G-G-U-
2	0.0	-U-U-U- • • • -A-G-A-
2	0.0	-U-U-U- • • • -G-G-A-
2	0.0	-C-U-U- • • • -C-G-A-
1	0.0	-G-G-U- • • • -C-U-A-
2	0.0	-U-G-U- • • • -G-C-G-
1	0.0	-U-C-G- • • • -G-G-U-
4	0.0	-G-A-G- • • • -U-U-U-
2	0.0	-G-G-A- • • • -U-U-U-
2	0.0	-G-G-G- • • • -U-C-U-
3	0.0	-U-U-U- • • • -G-A-G-
1	0.0	-U-G-G- • • • -A-U-C-
5	0.0	-C-C-U- • • • -U-C-G-
2	0.0	-G-G-G- • • • -C-U-C-

Anzahl	ΔG	Sequenz
1	0.0	-U-G-G- • • • -G-C-U-
2	0.0	-C-G-G- • • • -G-U-C-
1	0.0	-G-G-A- • • • -C-U-U-
9	0.0	-G-G-G- • • • -U-U-U-
3	0.0	-G-G-G- • • • -C-U-U-
5	0.0	-G-G-G- • • • -U-U-C-
1	0.0	-A-U-U-U- • • • • -U-G-G-A-
2	0.0	-U-U-U-U- • • • • -A-G-A-G-
1	0.0	-G-U-C-G- • • • • -C-G-G-U-
1	0.0	-G-G-U-C- • • • • -U-C-G-G-
1	0.0	-G-G-G-G- • • • • -U-U-U-C-
1	0.0	-G-G-U-G- • • • • -U-C-G-C-
1	0.0	-G-G-G-G- • • • • -U-U-U-U-
1	0.0	-U-G-G-G- • • • • -A-U-U-U-
1	0.0	-G-G-G-A- • • • • -U-U-U-U-
1	0.0	-G-G-U-G-G- • • • • • -U-C-G-C-U-
1	0.0	-G-G-U-G-G- • • • • • -C-U-A-U-C-
1	0.0	-G-G-G-G-U- • • • • • -U-U-U-U-A-

Anzahl	ΔG	Sequenz
1	0.0	-U-G-G-G-A- • • • • • -A-U-U-U-U-
1	0.0	-U-G-G-G-A- • • • • • -G-C-U-U-U-
1	-0.3	-U-G-U- • • • -A-U-G-
1	-0.3	-G-U-C- • • • -U-G-G-
3	-0.3	-A-U-G- • • • -U-G-U-
1	-0.3	-C-U-G- • • • -G-G-U-
3	-0.3	-U-G-U- • • • -G-U-A-
1	-0.3	-G-G-U- • • • -C-U-G-
1	-0.3	-C-U-G-U- • • • • -G-G-U-A-
1	-0.3	-G-G-G-G-U- • • • • • -U-U-C-U-G-
1	-0.6	-G-U-G-G- • • • • -U-G-U-C-
1	-0.6	-G-G-G-U-G- • • • • -U-C-U-G-U-
11	-1.2	-A-A- • • -U-U-
8	-1.2	-U-U- • • -A-A-
5	-1.8	-U-A- • • -A-U-
3	-1.8	-A-U- • • -U-A-
2	-1.8	-U-U-A- • • • -G-A-U-
2	-1.8	-A-U-G- • • • -U-A-U-
2	-1.8	-A-U-G- • • • -U-A-U-
3	-2.2	-A-G- • • -U-C-
3	-2.2	-G-U- • • -C-A-
4	-2.2	-U-G- • • -A-C-
5	-2.2	-A-C- • • -U-G-
6	-2.2	-C-A- • • -G-U-
12	-2.2	-C-U- • • -G-A-
6	-2.2	-U-C- • • -A-G-
5	-2.2	-G-A- • • -C-U-
1	-2.2	-A-C-G- • • • -U-G-U-
1	-2.2	-U-C-A- • • • -G-G-U-
1	-2.2	-G-A-G- • • • -U-U-C-
2	-2.2	-C-U-G- • • • -G-A-U-
2	-2.2	-U-U-C- • • • -G-A-G-
4	-2.2	-C-U-U- • • • -G-A-G-
1	-2.2	-A-C-U-U- • • • • -U-G-G-A-
1	-2.2	-G-G-G-U- • • • • -U-U-C-A-
1	-2.2	-U-G-G-G- • • • • -A-C-U-C-
1	-2.2	-C-U-G-C-U- • • • • • -G-A-U-G-G-
1	-2.2	-G-G-G-G-U- • • • • • -C-U-U-C-A-

Anzahl	ΔG	Sequenz
1	-2.5	-U-G-U-C-U- -A-U-G-G-A-
1	-2.5	-C-U-G-U-C- -G-A-U-G-G-
1	-3.0	-U-A-A- -A-U-U-
1	-3.0	-U-U-A-A- -G-A-U-U-
3	-3.2	-C-G- -G-C-
1	-3.2	-C-G-G- -G-C-U-
1	-4.0	-C-U-A- -G-A-U-
2	-4.0	-A-U-G- -U-A-C-
1	-4.0	-C-A-U- -G-U-A-
1	-4.0	-C-U-A-U- -G-A-U-G-
1	-4.4	-C-U-C- -G-A-G-
1	-4.4	-G-U-C- -C-A-G-
1	-4.4	-G-A-C- -C-U-G-
1	-4.4	-U-G-U-C- -G-C-A-G-
1	-4.4	-G-G-G-A-C- -U-U-C-U-G-
9	-5.0	-G-C- -C-G-
7	-5.0	-C-C- -G-G-
12	-5.0	-G-G- -C-C-
1	-5.0	-U-G-C- -G-C-G-

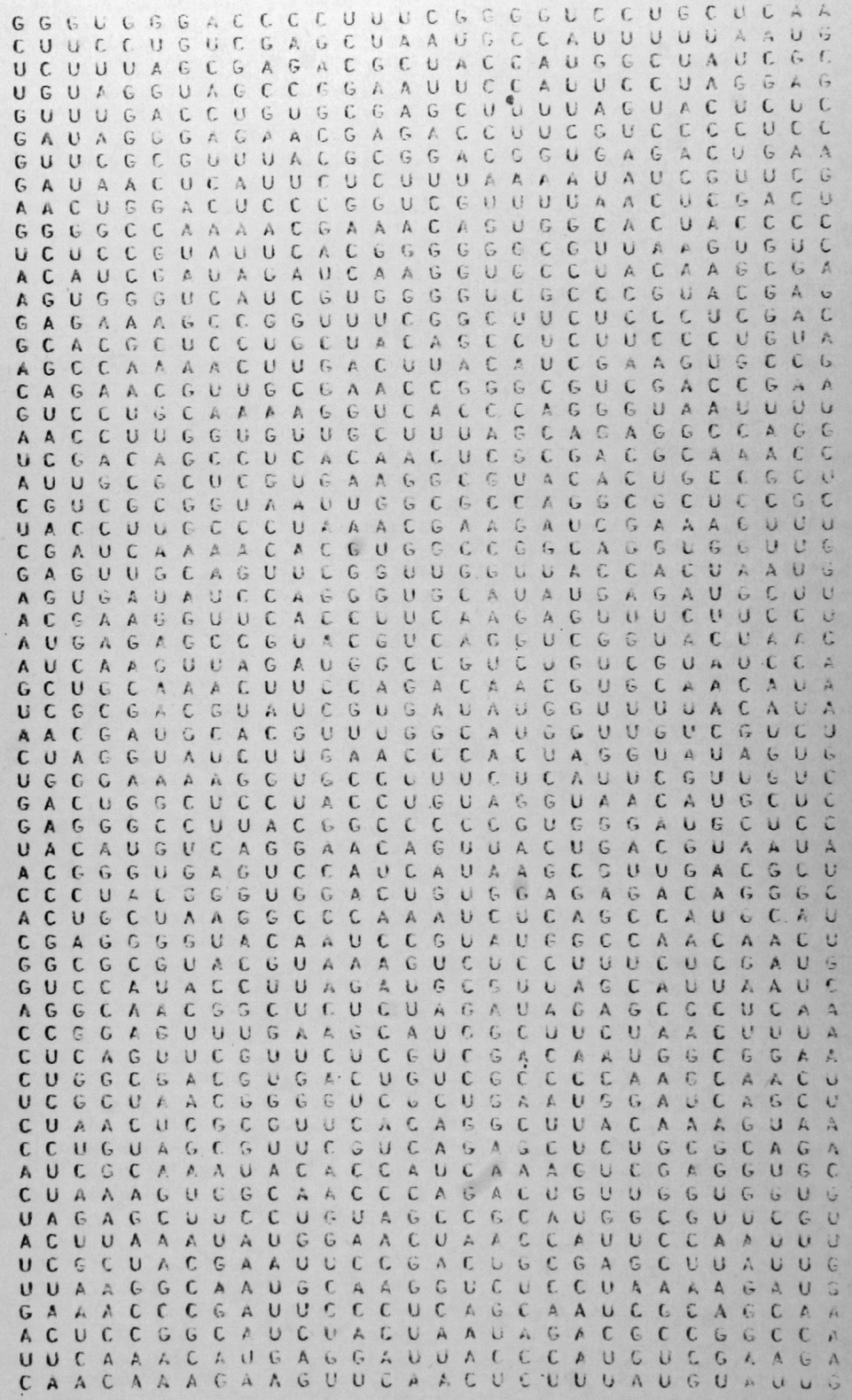
Anzahl	ΔG	Sequenz
2	-5.0	-G-C-U- -C-G-G-
1	-5.0	-G-G-U- -C-C-G-
1	-5.0	-C-C-U- -G-G-G-
2	-5.0	-G-G-G- -C-C-U-
5	-5.0	-G-G-G- -U-C-C-
1	-5.0	-U-G-C-C- -A-U-G-G-
1	-5.0	-G-G-G-A- -C-C-U-U-
1	-5.0	-G-G-G-G- -C-C-U-U-
1	-5.0	-C-G-G-G-G- -G-U-C-C-U-
1	-5.3	-G-G-U-G- -C-C-G-U-
1	-5.3	-G-G-G-U- -C-C-U-G-
1	-5.3	-A-U-C-G-C- -U-G-U-C-G-
1	-5.4	-C-G-A- -G-C-U-
1	-5.4	-U-C-G- -A-G-C-
1	-5.4	-U-U-U-C-G- -A-G-A-G-C-
1	-5.4	-U-U-C-G-G- -G-A-G-C-U-
1	-6.2	-C-U-A-C- -G-A-U-G-
1	-6.6	-U-G-U-C-U-U- -G-C-A-G-A-G-
1	-7.2	-G-C-U- -C-G-A-

Anzahl	ΔG	Sequenz
2	-7.2	-A-G-C- ····· -U-C-G-
1	-7.2	-C-C-U- ····· -G-G-A-
1	-7.2	-A-C-C- ····· -U-G-G-
1	-7.2	-G-G-U- ····· -C-C-A-
1	-7.2	-G-C-U-G- ····· -C-G-A-U-
1	-7.2	-U-G-C-U- ····· -G-C-G-A-
1	-7.2	-G-U-G-G- ····· -U-A-C-C-
1	-7.2	-A-G-C-U-A- ····· -U-C-G-G-U-
1	-7.2	-C-C-U-G-C- ····· -G-G-A-U-G-
1	-7.2	-G-G-U-G-G- ····· -C-C-A-U-C-
1	-7.2	-G-G-G-U-G-G- ····· -U-U-U-A-C-C-
1	-7.5	-C-C-U-G-U- ····· -G-G-A-U-G-
1	-7.8	-G-G-U-G-G-G-A- ····· -C-U-G-U-C-C-U-

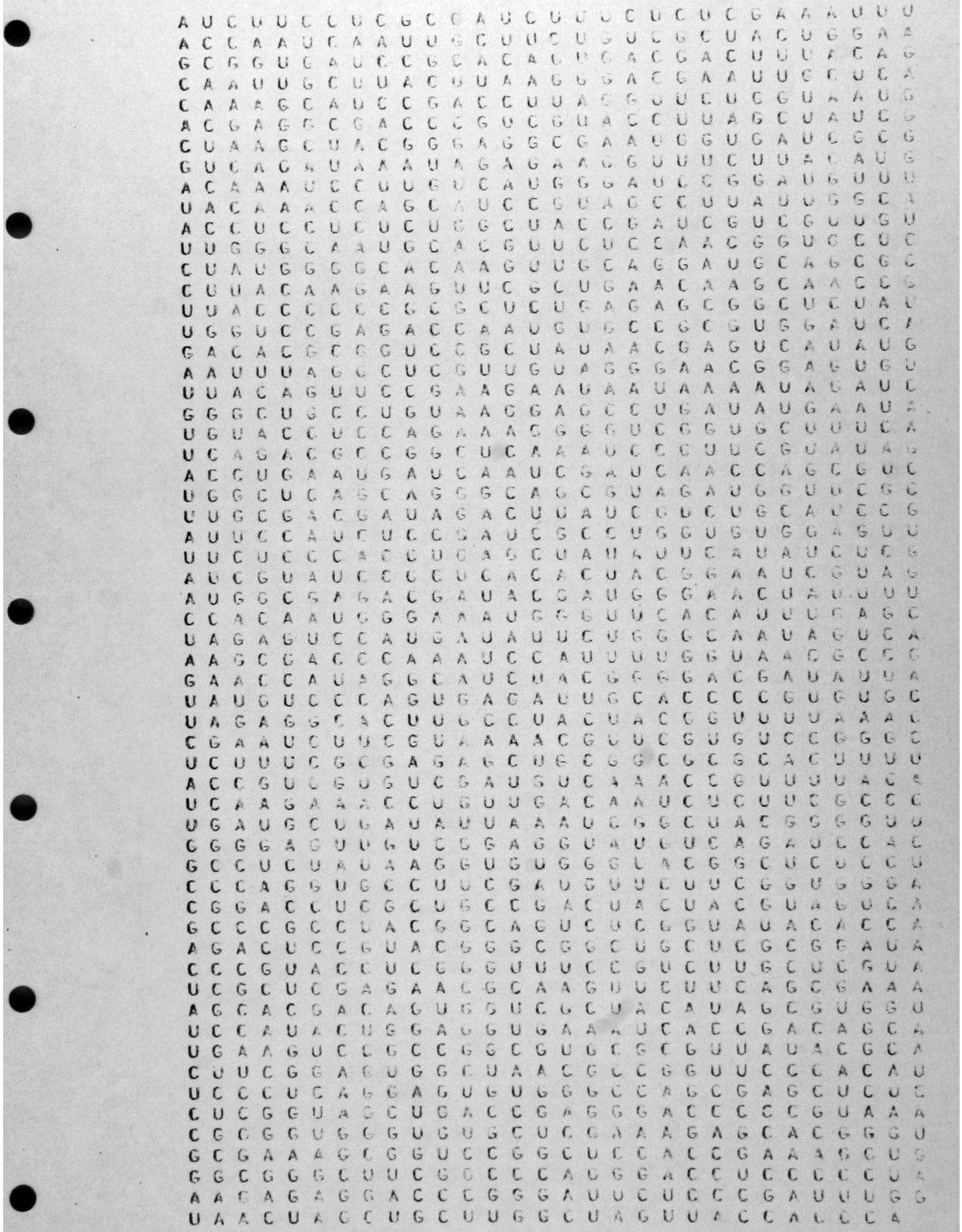
Anzahl	ΔG	Sequenz
1	-9.0	-U-A-G-C- ····· -A-U-C-G-
1	-9.0	-G-C-U-A- ····· -C-G-A-U-
1	-9.0	-G-C-U-A-U-C- ····· -C-G-A-U-G-G-
1	-9.4	-G-C-U-C- ····· -C-G-A-G-
2	-10.0	-C-C-C- ····· -G-G-G-
2	-10.0	-G-G-G- ····· -C-C-C-
1	-10.2	-G-C-U-A-A- ····· -C-G-A-U-U-
1	-12.2	-U-A-G-C-G- ····· -A-U-C-G-C-
1	-12.6	-U-A-G-C-G-A- ····· -G-U-C-G-C-U-
1	-14.0	-G-C-C-A-U- ····· -C-G-G-U-A-
1	-16.2	-G-C-U-A-C-C- ····· -C-G-A-U-G-G-
1	-24.7	-G-G-G-U-G-G-G-A-C-C-C-C- ····· -C-U-C-G-U-C-C-U-G-G-G-G-

Nucleotidsequenz des Bakteriophagen MS2

Leserichtung: von links nach rechts, zeilenweise
entspricht der 5'- 3'-Richtung



1.Fortsetzung:Sequenz MS2



Nucleotidsequenz des Bakteriophagen φX174

Leserichtung: von links nach rechts, zeilenweise
entspricht der 5'- 3'-Richtung

GAGUUUUAUCGCUUCCAUGACGCAGAACUU
AACAACUUUUCGGAUAAUUUCUGAUGAGUC
AAAAUUAUCUUGAUAAAGCAGGAAUUACUAC
UGCUUGUUUACGAAUUAAAUCGAAGUGGAC
UGCUGGCGGAAAAUGAGAAAAAUUCGACCUA
UCCUUGCGCAGCUCGAGAAGCUCCUUACUUU
GCGACCUUUCGCCAUCAACUAACGAUUCUC
UCAAAAAACUGACGCGUUGGAUGAGGAGAAG
UGGCUUAAUAUGCUUGGCACGUUCGUCAAG
GACUGGUUUAGAUAUGAGUCACAUUUUGUA
CAUGGUAGAGAUUCCUCUUGUUGACAUUUA
AAAGAGCGUGGAUUACUAUCUGAGAAAUCA
GCUGUUCAACCACUAAUAGGUAAGAAAUCA
UGAGUCAAGUUACUGAACAAUCCGUACGUU
UCCAGACCCCUUUGGCCUCCUCAUUAGCA
UUCAGGCUUCUGCCGUUUUGGAUUUAACCG
AAGAUGAUUUCGAUUUUCUGACGAGUAACA
AAGUUUGGAUUCCUACUGCCUCUCUGCGU
CUCGUCGCUGGCGUUGAGGCUUGCGUUA
GUACGCUGGACUUUGUAGGAUACCCUCGCU
UUCCUGCUCCUGUUGAGUUAUUAUUGCUC
UCAUUGCUUAUUAUGUUCAUCCCGUCAACA
UUCAAACGGCCUCUCUCAUCAUGGAAGGCG
CUGAAUUUACGGAAAACAUUAUUAAUGGCG
UCGAGCGUCCGGUAAAGCCGCUCAAUUGU
UCGCGUUUACCCUUGCGUGUACGCGCAGAA
ACACUGACCUUCCUUACCUGAACGAAAAA
ACGUGCGUCAAAAAUUACGCUCGGAAGCAA
UGAUGUAAUGUCUAAAGGUAAAACGAUUC
UGGCGCUCCGCCCUGGUCGUCCGCAGCCGUU
GCGAGGUACUAAAGGCAAGCGUAAAAGGCGC
UCGUCUUUGGUAUGUAGGUGGUCAACAAUU
UUAAUGCAGGGGCUUCGGCCCCUACUUGG
AGGAUAAAUUAUGUCUAAUAUUCAAACUGG
CGCGAGCGUAUGCCGCAUGACCUUUCCCA
UCUUGGCCUUCCUUGGCUGGUCAGAUGGUC
UCUUAUUACCAUUUCAACUACUCCGGUUAU
CGCUGGCGACUCCUUCGAGAUGGACGCCGU
UGGCGCUCUCCGUCUUCGAUUGCGUCG
UGGCCUUGCUAUUGACUCUACUGUAGALAU
UUUUACUUUUUAUGUCCCUCAUCGUCACGU
UUAUGGUCAACAGUGGAUUAAGUUCAUGAA
GGAUGGGUUAAUGCCACUCCUCCUCCCGAC
UGUUAACCAAACUACUGGUUAUUGACCA
UGCCGCUUUUCUUGGCACGAUUAACCCUGA
UACCAAUAAAAAUCCCCUAAAGCAUUUCUUUCA
GGGUUAUUUGAUACUAUAGCCUAUAGGUAA
AGCGCCGUGGAUGCCUGACCGUACAGGC
UAACCCUAAUGAGUUUUCUUCAUGGAGCU
UCGUUAUGGUUCUGCCUGCCAUCCUCAA
AAACAUUUGGACCUCCUCCGCUUUCCUGA
GACUGAGCUUUCUCGCCAAAUGACGACUC
UACCACAUCUAUUGACAUUAUGGGUCUGCA
AGCUGCUUAUGCUAACUUGCAUACUGACCA
AGAACGUGAUUACUUCAUGCAGCGUACCA
UGAGUUAUUCUUCAUUGGAGGUAAAACC
UCAUAUGACGCUGACAACCGUCCUUACUU
GUCAUGCCUCUAAUCUCUGGGCAUCUGSGC
UAUGAUCUUGAUGGAACUGACCAAACGUCG
UUAGGCCAGUUUUCUGCUCGUGUUCAAGG

1. Fortsetzung: Sequenz ϕX174

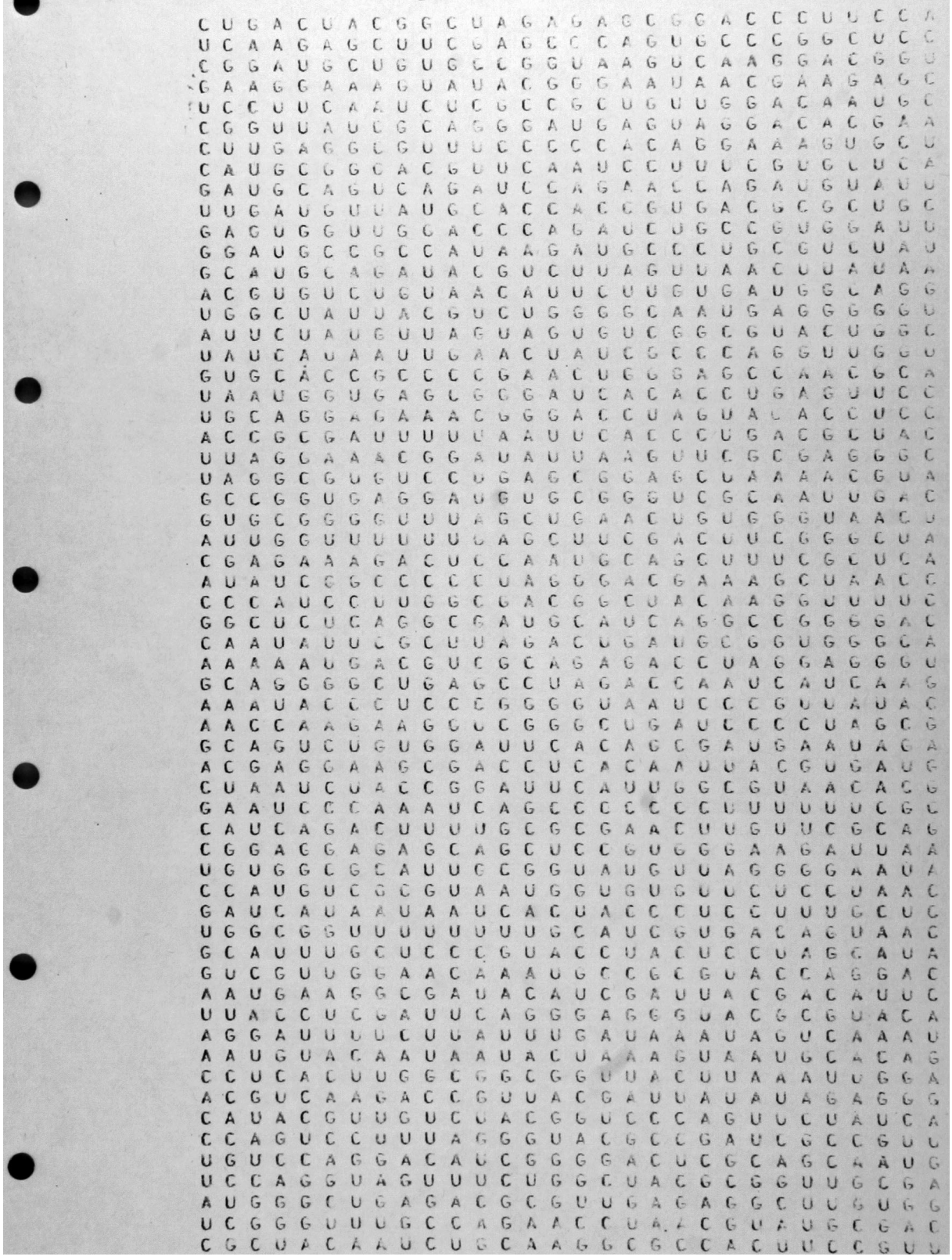
```
A C C U A U A A A C A U U C U G U G C C G C G U U U C U U U
G U U C C U G A G C A U G G C A C U A U G U U U A C U C U U
G C G C U G G U U C G U U U U C C G C C U A C U G C G A C U
A A A G A G A U U C A G U A C C U U A A C G C U A A A G G U
G C U U U G A C U U A U A C C G A U A U U G C U G G C G A C
C C U G U U U U G U A U G G C A A C U U G C C G C C G C G U
G A A A U U U C U A U G A A G G A U G U U U U C C G U U C U
G G U G A U U C G U C U A A G A A G U U U A A G A U U G C U
G A G G G U C A G U G G U A U C G U U A U G C G C C U U C G
U A U G U U U C U C C U G C U U A U C A C C U U C U U G A A
G G C U U C C C A U U C A U U C A G G A A C C G C C U U C U
G G U G A U U U G C A A G A A C G C G U A C U U A U U C G C
A A C C A U G A U U A U G A C C A G U G U U U C A G U C G U
U C A G U U G U U G C A G U G G A U A G U C U U A C C U C A
U G U G A C G U U U A U C G C A A U C U G C C G A C C A C U
C G C G A U U C A A U C A U G A C U U C G U G A U A A A A G
A U U G A G U G U G A G G U U A J A A C C G A A G C G G J A
A A A A U U U U A A U U U U U G C C G C U G A G G G G U U G
A C C A A G C G A A G C G C G G J A G G U U U U C J G C U U
A G G A G U U U A A U C A U G U U U C A G A C U U J U A U U
U C U C G C C A C A A U U C A A A C U U U U U U U C U G A U
A A G C U G G U U C U C A C U U C U G U U A C U C C A G C U
U C U U C G G C A C C U G U U U U A C A G A C A C C U A A A
G C U A C A U C G U C A A C G U J A U A U U U U G A U A G U
U U G A C G G U U A A U G C U G G U A A U G G U G G U J U U
C U U C A U U G C A U U C A G A U G G A U A C A U C U G U C
A A C G C C G C U A A U C A G G U U G U U U C A G J U G G U
G C U G A U A U U G C U U U U G A U G C C G A C C C U A A A
U U U U U U G C C C U G U U U G G J J C G C U U U G A G U C U
U C U U C G G U U C C C G A C U A C C C U C C C G A C U G C C
U A U G A U G U U U A U C C U U U G G A U G G U C G C C A U
G A U G G U G G U U U A U U A U A C C G U C A A G G A C U G U
G U G A C U A U U G A C G U C C U U C C C C G U A C G C C C
G G C A A U A A C G U C U A C G J U G G U U U C A U G G U U
U G G U C U A A C U U U A C C G C U A C U A A A A U G C C G C
G G A U U G G U U U C G C U G A A U C A G G U U A U U A A A
G A G A U U A U U U G U C U C C A G C C A C U U A A G U G A
G G U G A U U U A U G U U U G G U G C U A U U G C U G G C G
G U A U U G C U U C U G C U C U J G C U G G U G G C G C C A
U G U C U A A A A U U G U U U U G G A G G C G G U C A A A A A A G
C C G C C U C C G G U G G C A U U C A A G G U G A U G U G C
U U G C U A A C C G A U A A A C A A U A C U G U A G G C A U G G
G U G A U G C U G G U A U U A A A A U C U G C C A U U C A A G
G C U C U A A A U G U U C C U A A A C C C U G A U G A G G C C G
C C C C U A G U U U U U G U U U C G J G U G C U A U U G C C U A
A A G C U G G G U A A A A G G A C U U C U U G A A G G U A C G U
U G C A G G C U G G G C A C U U C C J G C C G U U U C U G G A U A
A G U U G C U U G A U U U G G G U J G G A C U U G G G J G G C A
A G U C U G C C G C U G A U A A A A G G A A A A G G G A U A C U C
G U G A U U A U C C U U G C U G C J G C A U U U C C C U G A G C
U U A A U G C U U G G G G A G C G J G C U G G U G C U J G A U G
C U U C C U C U G C C U G G G U A U G G U U G A C G C C G G A U
U U G A G A A A U C A A A A A A G A G C U U A C U A A A A A U G C
A A C U G G G A C A A A U C A G A A A A G A G A U U G C C G A G A
U G C A A A A A U G A G A G A C U C A A A A A A G A G A G A U U G C U G G
G C A U U C A G U C G G C G A C U U C A C G C C A G A A U A
C G A A A G A C C A G G G U A U A A J G C A C A A A A A U G A G A
U G C U U G C U U A A U U C A C A G A A G G A G U C U A C U G
C U G C G U U G C G U C U A U U A U G G A A A A C A C C C A A
U C U U U C C C A A G C A A C A G C A G G U U U C C C G A G A U
```

2.Fortsetzung: Sequenz X174

```
U A U G C G C C A A A U G C U U A C U C A A G C U C A A A C
G G C U G G U C A G U A U U U U A C C A A U G A C C A A A U
C A A A G A A A A U G A C U C G C A A G G U U A G U G C U G A
G G U U G A C U U A G U U C A A U C A G C A A A A C G C A G A A
U C A G C G G U A U G G C U C U U C U C A U A A U U G G C G C
U A C U G C A A A G G A U A U U U J C U A A A U G U C G U C A C
U G A U G C U G C U U C U G G U G U G U G G U U G A U A U U U U
U C A U G G U A U U G A U A A A A G C U G U U G C C C G A U A C
U U G G A A C A A A U U U C U G G A A A A G A C G G U A A A A G C
U G A U G G U A U U G G C U C U C U A A A U U U G U C U A G G A A
A U A A C C G G U C A A G G A A U U G A A C A C C C U C C C C A A A U U
G U A U G U U U U U C A U G C C U C C A A A U C U U G G A G G
C U U U U U U U A U G G G U U C C G U U C U U A A U U A C C C U U C
U G A A U G U C A C G C U G A U U A U U U U G A C U U U G G A
G C G U A U C G A G G C U C U U A A A C C U G C U A A U U G G A
G G C U U G U U G G C A U U U C U A C U C U U U C U C U C A A A U C
C C C A A U G C U U G G G C U U U C C A U A A A G C A G A A U G G A
U A A C C G C A U C A A A G C U C J U G G A A G A G A A G A U U C U
G U C U U U U U C G U A U G C A G G G C G U U G A G U U C G A
U A A A U G G U G A U A A U G U A A U G U U G A C G G C C A U A A
G G C U G C U U C U G A A C G U U C C G U G U G A U G A G U U U G U
A U C U G U U U A C U G A G A A A G U U J A A A U G G A A U G A A A U U
G G C A C A A A U G C U A C A A A U G G U G C U C C C C C C A A A C U
U G A A U A U U U A A U A A A C A C U A A U A G A C C A C C G C C C
C G A A A G G G G A C G A A A A A A J G G U U U U U U A G A G A A A
C G A G A A G A A C G G G U U A C G C A G U U U U U G C A A A G C U
G G C U G C U G A A C G C C C J C J U A A A G G A A A U A U U C C G
C G A U G A G U A U A A A U U A C C C C C A A A A A A G A A A A G G
U A U U A A A G G A U G A G U G U U C A A G A A U U G C U G G G A
G G C C U C C A C U A A A G A U A A J C G C G U A G A G G C U U
U G C U A U U C A G C G U U U G A A U G A A A U G C C A A A U G C G
A C A G G C U C A A U G C U G A A U G G U U G G G U U U U A U C G U
U U U U G A C A C U C U C A C G J U G G C U G A C G A A C C G
A U U A G A G G C G U U U U A A U G A A U A A A U C C C A A A U G C
U U U G C G U G A C U A A U U U U C G U G A U A A U U G G U C G
U A U G G U U C C U U G C C U G C C C G A G G G U C G C A A G G C
U A A A U G A A U U C A C A C G C C G A C U G C U A A U C A G U A
U U U U U G U G U G U G C C U G A A G J A U G G G U A C A A G C U A A A
U G G C C G U C U U C C A A U U U C C C A U G C G G U G C A C U U
U A U G C G G A C A C U U C C C U A C A G G U A G C G U U U G A
C C C U A A A U U U U U G G U C G U C C G G G U A C G C A A A U C G
C C G C C A G U U A A A A U A G C J U G C A A A A A U A C G U G
G C C U U A A U G G G U U A A C A G U A A J G C C C A A U C G C A G U
U C G C U A C A C G C A G G A A C G C U U U U U U C A C G J U C
U G G G U U G G G U U G G U G G G C C C U G G U U G G A U G C U A A A A G G
U G A A G C C C G C U U A A A A A G C U A A C C A A G U U A A U A A U G G C
U G U U U G G G U U U U C U A A U G U G U G G C U A A A A U A C G U U A A A
C A A A A A A G U C A G A A U A A U G G A A C C U U G C U G C U A A A
A G G G U C U A A G G A A G C C U A A A A A G A A A U G G G A A C A A A C U C
A C U A A A A A A A C C C A A A G C U G U U C G C U A C U U C C C C A A
G A A A G C U G G U U C C A A G A A A U C C A G A A A U G A A G C C C G C A A
C U U U C G G G G A U G A A A A A A U G G C U C U C A C A A A U G A A C A A A
U C U G G U C C C A C G G G A A G U G C C J U A A A U C C C A A A C C U U A C
C A A A G C U G G G G U U A C G A A C G C C G A A C G C C C G J U C C A A
C C C A G A U A A U U G G A A A G C C A A G A A A C G C C A A A A A A A G A G A A
G A U G A G A A U U G A A G G C U G G G G A A A A A G U U A C U G G U
A G C C G A C G U U U U U G G C G G C G C G C A A C C U G G U G A C
G A C A A A A U C C U G C U C A A A A U U U A U G C G C G C U U C
G A U A A A A A A U G A A U U G G C G U A U C C C A A C C C U G C A
```

Nucleotidsequenz des Zufallsmessengers

Leserichtung: von links nach rechts, zeilenweise
entspricht der 5'- 3'-Richtung



1.Fortsetzung: Sequenz Zufallsmessenger

2.Fortsetzung: Sequenz Zufallsmessenger

```
G U C U G C U C U A G A U C C G A U C C C G U U A A U A A G
G G G C A A C G C A U C U C G U C U C U G A G C A U U U C G
U C C G C A C A G G A A G A C A A U U G A A A U U U G A A C
U A U G U A U C G A G G U A C C C C U G C U C C C U A U C U
A A A G C A G C G A G G U C U U C C A A U U C G A A U A A G
A G G G G G C U U A G U A U U G A A A G A C A U A A G A A U
G U U C A U U C G C U U G A A G A A U A C A A A C A C G G G
U A U U A A U G U G U A A G G G U G U C A U U G A A C G U A
U C U G U U C A C A A A U U C C C L G G G U U U C A A A U C A
G G G C C U A G C U G A U A C A G G G U G U A U C U G G U G
C U G G U C U A A U A C U U C G G G U G U C U A C C U G G C
A A U C G G U G U G U C G G A U G U U A A A U A A G G G A A
A U U A A G A C A C A C U A U C U C G A U G A A U U G C A U
G A G U U A C U C A A G C U U A U G A C G C G A C U G U C C
U A C U C G A U U G G A C G G G U G G A C A U U U A U C C C
A A A C A A G A U C A G U U A U A G A C C U U A G C G A A A
G C A U G C A U C C U C G U C C G C A U C C C C C U U C G G
A U A G G U G A A C U C A G A C A G G G A A G A A C C U G U
G C U C C G G A C U G G C U G A U G G G G C U A U U G U G A
C A A U A A C C C U C G A A G C C A G C G G G G U G A G A C
U A U U C C G A G G U G A C U U G G U U C A G C U C U U G A
C C A G G A C G C G C U G U A G C C G A U G C C C U C A A G
C U U C C G G A A U C A U C C C A G A G A G G U G U G U U C G A C
C G A C G A A U C G G C C G G U U C G C C A U C A U A C A C G
U U A C A C C C G G A A U G U A C C G G A A U U U A U A U G A
A A A C A C A U U A U C C U A U A C G G C G C G U A U C C G
G G U A U A A A U U U U G A C C C C C U C A A C U G U C C
G C U C A A U A C G C C U C G A U G C G C A A C A U C C G A
G C A A A C C G A G C A G C A A A U U G A G G C G A C G U U
C G G U U U A A C C U U U G A G A G C U U A G G C U U C C G
G G C A U G A U A A C C U C U U G U U C C U U C A G G U U C
A A G G A G A C A U U A U C U G G G A U G A A G G U C C U C
C A G G C A C G A G U U C U U U G U G C U A C G G A G U G C
C A G A A C A A C G G A U G C C G C U U A A A C C G C A A U
A G A C U A G C A G A U G C A A A C U G C C G G A U U U A U
C C C C C A U U G U G C A A C C G C G U A U G U G C A U A C
A C G G U G U A A A A G U G A G C C C C G G C A C U U C G G
A U A U G U C U U U G G U A U G U G C G C C G C U A U A C U
A U U A C A A U C G U G A A U A G A A U U C A C A U U A U U U
G U G C A G C A C C C U U G G U C U A G A U G U A C A U G C
A A A G U U U A A A U G C C U C A C C U G C C G A U G C G G G G
A A U G U G U U U U U G C A U A G G C C A A G G G U U A A C A
A A C U G G C U A C U G U G A G G C U C C C U C G A G U C U G
A U A U A U U C U C U A C C U G C C G A G G G A G C A U C C
C C G U U U A A A U C G C A G U A U G U C G C G U C C C U C C
C A A A U A U G C C U A A U C A U C C C U G C C C C U U C A
A G G G C A U C G U U A C C G U G U G U
```

7 LITERATURVERZEICHNIS

(1) ADAMS J.M.,JEPPESEN P.G.N,SANGER F., BARREL B.G. "A nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA"
Nature 223,S.1009 (1969)

(1a) ASHBY R.W. "Einführung in die Kybernetik"
Frankfurt,(1974) (Buch)

(2) BALL L.A. "Implications of Secondary Structure in mRNA"
Journal of theoretical Biology 36, S.313-320 (1972)

(3) BARREL B.G.,AIR G.M.,HUTCHISON C.A. "Overlapping genes in bacteriophage ϕX174"
Nature 264,S.34-41 (1976)

(4) BORER P.N.,DENGLER B.,TINOCO I.,UHLENBECK O.C. "Stability of RNA Double-stranded Helices"
Journal of Molecular Biology 86(4),S.843 (1974)

(5) BROWNLEE G.G. "Determination of sequences in RNA"
Amsterdam (1972) (Buch)

(6) BURTON K.,LUNT M.R.,PETERSON G.B., SIEBKE J.C.
CSHS 28,S.27 (1963)

(7) BURTON K.,PETERSON G.B.
Biochemical Journal 75, S.17-27 (1960)

(8) CONTRERAS R.,VANDENBERGHE A.,VOLCHAERT G., MIN JOU W.,FIERS W. "Studies on the Bacteriophage MS2. Some nucleotide sequences from the RNA-Polymerase Gene"
FEBS Letters 24,S.339-342 (1972)

(9) DeWACHTER R., MERREGAERT J.,VANDENBERGHE A.,CONTRERAS R.,FIERS W. "Studies on the Bacteriophage MS2. The Untranslated 5'-Terminal Nucleotide Sequence Preceding the first Cistron"

European Journal of Biochemistry 22, S.400-414 (1971)

(10) FIERS W.,CONTRERAS R.,DeWACHTER R.,STAEGEMAN G.,MERREGAERT J., MIN JOU W.,VANDENBERGHE A. "Recent progress in sequence determination of bacteriopage MS2 RNA" Biochimie 53,S.495-506 (1971)

(11) FIERS W.,CONTRERAS R.,DUERINCK F.,HAEGEMAN G.,MERREGAERT J.,MIN JOU W.,RAEYMAKERS A.,VOLCHAERT G.,YSEBAERT M. "A-Protein gene of bacteriophage MS2" Nature 256(5515),S.273-278 (1975)

(12) FIERS W.,CONTRERAS R.,DUERINCK F.,HAEGEMAN G.,ISERENTANT D.,MERREGAERT J.,MIN JOU W.,MOLEMANS F.,RAEYMAEKERS A.,VANDENBERGHE A.,VOLCHAERT G.,YSEBAERT M. "Complete nucleotide sequence of bacteriophage MS2 RNA:primary and secondary structure of the replicase gene" Nature 260(5551),S.500-507 (1976)

(13) FIGUEROA R.,SOTO A.,GONZALEZ G.,PIEBER M., ROMERO C.,TOHA J.C. "Base Sequence of an Non-self-complementary Messenger RNA of Human Heart Cytochrome c" Journal of theoretical Biology 36,S.321-326 (1972)

(14) GALIBERT F.,SEDAT J.W.,ZIFF E.B. "Direct determination of DNA Nucleotidesequences. Structure of a fragment of Bacteriophage φX174 DNA Journal of molecular Biology 87,S.377-407 (1974)

(15) GRALLA J.,DeLISI C. "mRNA is expected to form stable secondary structures" Nature 248(5446) S.330-332 (1974)

(16) VON HEIJNE F.,NILSSON L.,BLOMBERG C. "Translation and mRNA secondary structure" Journal of theoretical Biology 68(3), S.321 (1977)

(17) HOLLEY R.W.,APGAR J.,EVERETT G.A.,MADISON J.T.,MARQUISE M.,MERRILL S.H.,PENSWICH J.R., ZAMIR A.
Science 147,S.1462 (1965)

(18) JACK A.,LADNER J.E.,KLUG A. "Crystallographic Refinement if yeast Phe t-RNA at 2.5 Å Resolution"
Journal of molecular Biology 108(4),S.619-649 (1976)

(19) JACOBSON A.B. "Studies on secondary structure of single-stranded RNA from bacteriophage MS2 by electron microskopy"
Proceedings of the National Academy of Sciences USA 73(2), S.307-311 (1976)

(20) JAY E.,ROYCHOUD.R., WU R. "Nucleotide-sequence with elements of an unusual 2-fold rotational symmetry in region of origin of replication of SV40 DNA"
Bioc.Biop.R. 69(3),S.678-686 (1976)

(21) JORDAN B.R. "Computer Generation of Pairing Schemes for RNA Molecules"
Journal of theoretical Biology 34,S.363-378 (1971)

(22) KIM S.H. "Three-dimensional strcture of tRNA"
Progress in Nucleic Acis Research and Molecular Biology 17 (1976)

(23) KIM S.H.,QUIGLEY G., SUDDATH F.L., Mc PHERSON A.,SNEDEN D.,KIM J.J.,WEINZIERL J.,BLATTMANN P.,RICH A. "The Three-Dimensional Structure of Yeast Phenylalanine Transfer RNA: Shape of the Molecula at 5.5-Å Resolution"
Proceedings of the National Academy of Sciences USA 69, S.3746-3750 (1972)

(24) KLÄMBT D. "A Model for Messenger RNA Sequences Maximizing Secondary Structure Due to Code Degeneracy"
Journal of theoretical Biology 52,S.57-65 (1975)

(25) KLÄMBT D.,RICHTER O. "Computer Programs

for Studying Conformations in Ribonucleic Acids"
Journal of theoretical Biology 58,S.319-324 (1976)

(25a) KORNBERG A. "DNA Synthesis"
San Francisco (1974) (Buch)

(26) LAPIDUS I.R.,ROSEN B.,HEPPERLE R. "Secondary Structure of Qβ RNA"
Journal of theoretical Biology 64,S.587-595 (1977)

(27) LAUX B.,DENNIS D.,WHITE H.B. "Human α-Chain Globin Messenger: Prediction of a Nucleotide Sequence"
Biochemical and Biophysical Research Communications 54,S.894-898 (1973)

(28) LESK A.M. "A Combinatorial Study of the Effects of Admitting Non-Watson-Crick Base Pairings and of Base Composition on the Helix-forming Potential of Polynucleotides of Random Sequence"
Journal of theoretical Biology 44,S.7-17 (1974)

(29) McMAHON J.E., PIPAS J.M. "Predicting RNA Secondary interactions from primary sequence"
International Journal of quantum Chemistry (Symposium-2) S.129-131 (1975)

(30) MIN JOU W., MERREGAERT J.,CONTRERAS R., DUERINCK F.,HAEGEMAN G.,RAEYMAEKERS A., VOLCKAERT G.,YSEBAERT M.,FIERS W. "Bacteriophage-MS2 RNA. Nucleotide-Sequence Determination of A-Protein Gene"
H-S. Z.Physl. 355(10) S.1231 (1974)

(31) MIN JOU W.,HAEGEMAN.,YSEBAERT M.,FIERS W. "Nucleotide Sequence of Gene Coding for the Bacteriophage MS2 Coat Protein"
Nature 237,S.82-88 (1972)

(32) PIPAS J.M., McMAHON J.E. "Method for Predicting RNA Secondary Structure"
Proceedings of the National Academy

of Sciences USA 72,S.2017-2021 (1975)

(33) SANGER F.,AIR G.M.,BARRELL B.G.,BROWN N.L.,COULSON A.R.,FIDDES J.C.,HUTCHISON C.A.,SLOCOMBE P.M.,SMITH M. "Nucleotide Sequence of Bacteriophage φX174 DNA
Nature 265,S.677-695 (1977)

(34) SANGER F.,DONELSON J.E., COULSON A.R., KÖSSEL H.,FISCHER D. "Use of DNA Polymerase I primed by a synthetic Oligonucleotide to determine a nucleotide sequence in Phage FL DNA"
Proceedings of the National Academy of Sciences USA 70,S.1209-1213 (1973)

(35) SCHOTT H. "(GE) Chemical Syntheses of a Phage-specific DNA fragment"
Makromolecular Chemistry 175,S.1683-1693 (1974)

(36) SEKIYA T.,GAIT M.J.,NORIS K.,RAMANOOR.B., KHORANA H.G. "Studies on Polynucleotides. 144. Nucleotide-Sequence in Promoter Region of gene for an E.coli Tyr-tRNA"
Journal of biological Chemistry 251(15), S.4481-4489 (1976)

(36a) STEWART P.R.,LETHAM D.S. (ed.) "The Ribonucleic Acids"
New York (1973) (Buch)

(37) TINOCO I.,BORER P.N.,DENGLER B.,LEVINE M.D., UHLENBECK O.C.,CROTHERS D.M.,GRALLA J. "Improved Estimation of Secondary Structure in Ribonucleic Acids"
Nature New Biology 246,S.40-41 (1973)

(38) TINOCO I.,UHLENBECK O.C.,LEVINE M.D. "Estimation of Secondary Structure in Ribonucleic Acids"
Nature 230,S.362-367 (1971)

(39) VOLCKAERT G.,MIN JOU W.,FIERS W. "Analysis of P-32-labeled Bacteriophage MS2 RNA by a Mini-fingerprinting Procedure"
Analytical Biochemistry 72(1-2),S.433-446 (1976)

(40) Yockey H.P. "An Application of Information theory to the Central Dogma and the Sequence Hypothesis"
Journal of theoretical Biology 46,S.369-406 (1974)

(41) ZIFF E.B.,SEDAT J.W.,GALIBERT F. "Determination of Nucleotide-Sequence of a fragment of Bacteriophage φX174 DNA
Nature New Biology 241,S.34-37 (1973)

Erklärung:

Hiermit versichere ich,dass ich die Arbeit selbst verfasst und keine anderen Hilfsmittel als die angegebenen benutzt habe.
Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht. Dies gilt auch für Zeichnungen, Kartenskizzen,bildliche Darstellungen usw.